Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América
Facultad de Ingeniería de Sistemas e Informática
Escuela Profesional de Ingeniería de Sistemas

Sistema inteligente basado en redes neuronales, máquina de soporte vectorial y random forest para la predicción de deserción de clientes en microcréditos de bancos

TESIS

Para optar el Título Profesional de Ingeniero de Sistemas

AUTORES

Dennys AGUILAR VILCA

Julio Cesar CAMARGO RAMOS

ASESOR

Dr. David Santos MAURICIO SÁNCHEZ

Lima, Perú

2021

Referencia bibliográfica

Aguilar, D. & Camargo, J. (2021). *Sistema inteligente basado en redes neuronales, máquina de soporte vectorial y random forest para la predicción de deserción de clientes en microcréditos de bancos*. [Tesis de pregrado, Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería de Sistemas e Informática, Escuela Profesional de Ingeniero de Sistemas]. Repositorio institucional Cybertesis UNMSM.

Hoja de Metadatos complementarios

Código ORCID del autor	"—"
DNI o pasaporte del autor	Dennys Aguilar Vilca 71328761 Julio Camargo Ramos 70583630
Código ORCID del asesor	https://orcid.org/0000-0001-9262-626X
DNI o pasaporte del asesor	06445495
Grupo de investigación	NO
Agencia financiadora	NO
Ubicación geográfica donde se desarrolló la investigación	Lugar: Facultad de Ingeniería de Sistemas e Informática Coordenadas geográficas: -12.052996002294023, -77.08572910626616
Año ó rango de años en que se realizó la investigación	2017 - 2020
Disciplinas OCDE	Física atómica, molecular y química https://purl.org/pe-repo/ocde/ford#2.02.04 Sistemas de automatización, Sistemas de control https://purl.org/pe-repo/ocde/ford#2.02.03

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
FACULTAD DE INGENIERIA DE SISTEMAS E INFORMATICA
Escuela Profesional de Ingeniería de Sistemas

Acta Virtual de Sustentación de Tesis

Siendo las 16:00 horas del día 26 de marzo del año 2021 se reunieron virtualmente los docentes designados como miembros de Jurado de Tesis, presidido por el Mg. Hugo David Calderón Vilca (Presidente), Ing. Ana María Huayna Dueñas (Miembro) y el Dr. David Santos Mauricio Sánchez (Miembro Asesor), usando la plataforma Meet para la sustentación Virtual de la tesis Intitulada: **"SISTEMA INTELIGENTE BASADO EN REDES NEURONALES, MÁQUINA DE SOPORTE VECTORIAL Y RANDOM FOREST PARA LA PREDICCIÓN DE DESERCIÓN DE CLIENTES EN MICROCRÉDITOS DE BANCOS"**, del Bachiller: **Julio Cesar CAMARGO RAMOS;** para obtener el Título Profesional de Ingeniero de Sistemas.

Acto seguido de la exposición de la Tesis, el Presidente invitó al Bachiller a dar las respuestas a las preguntas establecidas por los Miembros del Jurado.

El Bachiller, en el curso de sus intervenciones demostró pleno dominio del tema, al responder con acierto y fluidez a las observaciones y preguntas formuladas por los señores miembros del Jurado.

Finalmente habiéndose efectuado la calificación correspondiente por los miembros del Jurado, el bachiller obtuvo la nota de diecisiete (17).

A continuación el Presidente del Jurado Mg. Hugo David Calderón Vilca, declara al Bachiller **Ingeniero de Sistemas.**

Siendo las 17:21 horas, se levantó la sesión.

Presidente
Mg. Hugo David Calderón Vilca

Miembro
Ing. Ana María Huayna Dueñas

Miembro Asesor
Dr. David Santos Mauricio Sánchez

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
FACULTAD DE INGENIERIA DE SISTEMAS E INFORMATICA
Escuela Profesional de Ingeniería de Sistemas

Acta Virtual de Sustentación de Tesis

Siendo las 16:00 horas del día 26 de marzo del año 2021 se reunieron virtualmente los docentes designados como miembros de Jurado de Tesis, presidido por el Mg. Hugo David Calderón Vilca (Presidente), Ing. Ana María Huayna Dueñas (Miembro) y el Dr. David Santos Mauricio Sánchez (Miembro Asesor), usando la plataforma Meet para la sustentación Virtual de la tesis Intitulada: **"SISTEMA INTELIGENTE BASADO EN REDES NEURONALES, MÁQUINA DE SOPORTE VECTORIAL Y RANDOM FOREST PARA LA PREDICCIÓN DE DESERCIÓN DE CLIENTES EN MICROCRÉDITOS DE BANCOS",** del Bachiller: **Dennys AGUILAR VILCA;** para obtener el Título Profesional de Ingeniero de Sistemas.

Acto seguido de la exposición de la Tesis, el Presidente invitó al Bachiller a dar las respuestas a las preguntas establecidas por los Miembros del Jurado.

El Bachiller, en el curso de sus intervenciones demostró pleno dominio del tema, al responder con acierto y fluidez a las observaciones y preguntas formuladas por los señores miembros del Jurado.

Finalmente habiéndose efectuado la calificación correspondiente por los miembros del Jurado, el bachiller obtuvo la nota de diecisiete (17).

A continuación el Presidente del Jurado Mg. Hugo David Calderón Vilca, declara al Bachiller **Ingeniero de Sistemas.**

Siendo las 17:21 horas, se levantó la sesión.

Presidente
Mg. Hugo David Calderón Vilca

Miembro
Ing. Ana María Huayna Dueñas

Miembro Asesor
Dr. David Santos Mauricio Sánchez

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

Agradecimientos

A nuestras familias por brindarnos la confianza, el apoyo y coraje para concluir exitosamente este trabajo.

A los diferentes colegas, amigos y profesores de la UNMSM por el apoyo brindado en la realización de este proyecto.

Al profesor Dr. David Mauricio Sánchez, por su orientación y dedicación para que este trabajo se cumpla con todos los objetivos planteados.

Y por encima de todo doy gracias a Dios.

Resumen

La deserción de clientes bancarios es un problema que afecta actualmente a las empresas de todos los sectores y en todos los países. Por su parte, el sector financiero es uno de los más importantes debido a la gran cantidad de clientes y dinero que estos aportan.

Las empresas invierten dinero para realizar un seguimiento a los clientes y poder identificar patrones que puedan evidenciar si un cliente va a dejar de hacer negocios con la empresa, pero muchas veces las maneras manuales de realizarlas presentan deficiencias de tiempo y de pérdida de dinero.

En la literatura es común ver modelos de predicción de deserción de clientes bancarios microcréditos, el punto débil de estos es que solo aplican una técnica para realizar propiamente la predicción. En virtud de esto, se propone un sistema inteligente basado en un modelo híbrido que combina tres técnicas para proporcionar mejor precisión que la observada en la literatura; estas son Máquinas de Soporte Vectorial, Redes Neuronales y Random Forest.

Los resultados numéricos obtenidos del experimento realizado a un banco peruano con un conjunto de datos de 24 420 clientes presentan una precisión de 97.38%, el cual mejora los resultados de la literatura.

Abstract

The dropout problem of bank customers is a problem that currently affects companies of all sectors and in all countries. The financial sector is one of the most important because of the large number of clients and money they bring.

Companies spend money to track customers and to identify patterns that can show whether a client will stop doing business with the company, but often perform manual ways deficient lost time and money.

In the literature is common to see prediction models dropout bank customers, the weakness of these models is that they only apply a technique to properly perform the prediction. Under this, we propose to use an Intelligence System based on a Hybrid Model, which combines the three techniques that provide better accuracy rate in the literature, these support vector machines, neural networks and random forest.

Numeric results that we obtained from the experiment, which was applied to a Peruvian Bank that has a data set of 24 420 clients presents an accuracy of 97.38% that result is better than the others do found in the state of art.

1 Capítulo 1. Introducción

1.1 Antecedentes

No existe una sola manera para definir la deserción de clientes, sin embargo, en líneas generales se define como la posibilidad de que un cliente deje de hacer negocios con una determinada empresa en un determinado tiempo, ya sea de manera voluntaria como involuntaria (V. Rabi, 2016). La deserción de clientes se ha convertido en un problema importante y es uno de los retos principales a los cuales se enfrentan muchas empresas de todo el mundo. Esta pérdida es una oportunidad para que los competidores puedan ganar un cliente mientras que la reducción de la deserción y la retención de los clientes actuales son la mejor estrategia de marketing costo-beneficio, debido a la gran competencia entre las compañías, estas necesitan enfocarse en retener los clientes actuales satisfaciendo sus necesidades de una manera efectiva (Farid Zhirazi, 2018).

En la actualidad las Micro y Pequeñas empresas (MYPES) en el Perú son de vital importancia para la economía de nuestro país. Según el Ministerio de Trabajo y Promoción del Empleo, tiene una gran significación por que aportan con un 40% al PBI, y con un 80% de la oferta laboral, sin contar con el autoempleo que genera (Sánchez Barraza, 2014). Por ello, resulta evidente que se requería elevar la productividad de la microempresa para lograr un incremento en el PBI del Perú. Entre las políticas relevantes para dicho fin destaca la promoción de los microcréditos.

En los últimos 15 años "se ha observado en la economía peruana, un importante dinamismo del sector financiero orientado a ofrecer servicios financieros a microempresas, empresas familiares o productores individuales" (Quispe, Leon, & Contreras, 2011).

Con este tipo de financiamiento, los microempresarios pueden incrementar su productividad a través de la inversión en activos para la producción. Se generarían así mayores ventas, ingresos familiares y puestos de trabajo. Asimismo, el crédito en pequeña escala puede beneficiar a las mujeres con iniciativa empresarial, contribuyendo a superar las barreras de género, tanto en el hogar como en el mercado laboral. (Aramburú & Portocarrero, 2002)

Hasta la fecha se han realizado numerosas investigaciones para la predicción de la deserción de clientes en la banca, sin embargo presentan dos puntos débiles notorios: La primera es que la gran mayoría de estas han sido realizadas en Europa y Asia, y a pesar que muchas de ellas

cuentan con un alto grado de acierto, no pueden ser del todo aceptados, debido a que las características de los clientes son muy diferentes. La segunda es que solo aplican una técnica para realizar propiamente la predicción, en virtud de esto, se propone un sistema inteligente basado en un modelo híbrido que combina tres técnicas para proporcionar mejor precisión en la literatura; estas son Máquinas de Soporte Vectorial, Redes Neuronales y Random Forest.

1.2 Problemática

El problema de predicción de la deserción de clientes en bancos trata acerca de identificar con anticipación cuál es el nivel de riesgo de deserción de un cliente micro-crédito en la banca de la ciudad de Lima.

1.3 Importancia

El problema de la deserción de clientes en microcréditos reviste importancia a nivel mundial, en todos los países las empresas financieras hacen frente a ello.

A Febrero del 2019 el sistema financiero peruano cuenta con 16 bancos privados, 4 entidades financieras estatales, 11 empresas financieras, así como 16 de las llamadas instituciones microfinancieras no bancarias, que incluyen 6 Cajas Rurales de Ahorro y Crédito (CRAC), 12 Cajas Municipales de Ahorro y Crédito (CMAC) y 9 Empresas de Desarrollo de la Pequeña y Microempresa (EDPYME). Es decir, hay una alta diversificación institucional en el mercado financiero regulado peruano. Este contexto es resultado del gran dinamismo financiero observado durante la década reciente, tal como se indicó antes. Estas empresas financieras, que se muestran en la Figura 1-1, prestan servicios de microcrédito.

Empresas de Operaciones Múltiples	Activos a Febrero 2019		
	Número de Empresas	Monto (S/ Millones)	Participación (%)
Banca Múltiple	16	389 023	89,35
Empresas financieras	11	14 948	3,43
Cajas municipales (CM)	12	26 991	6,20
Cajas rurales de ahorro y crédito (CRAC)	6	1 898	0,44
Entidades de desarrollo de la pequeña y microempresa (Edpyme)	9	2 514	0,58
TOTAL	**54**	**435 374**	**100**

Figura 1-1. Entidades financieras que brindar servicios de microcréditos

Fuente: elaboración propia con base en SBS (s.f.)

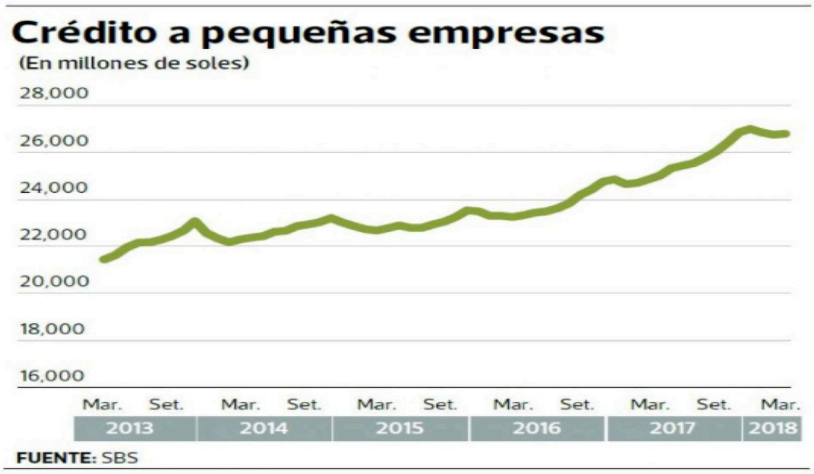


Figura 1-2. Montos otorgados por microcréditos a pequeñas empresas hasta marzo de 2018

Fuente: elaboración propia con base en SBS (s.f.)

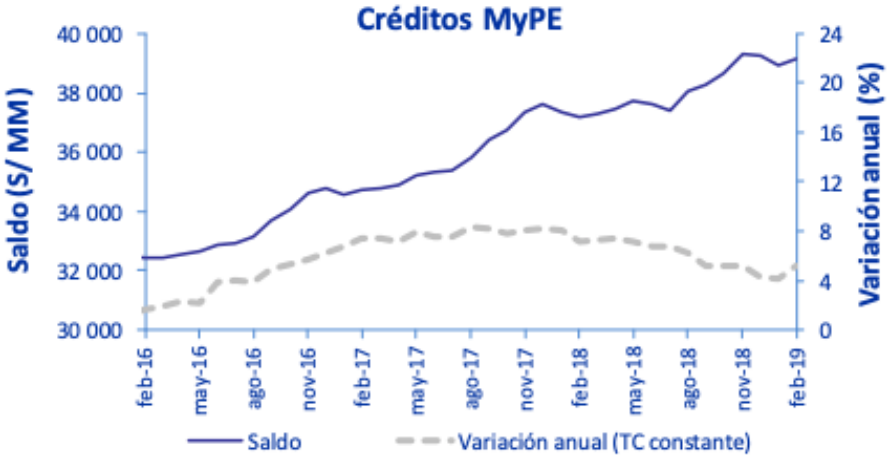
Entre los microcréditos se encuentran una serie de créditos destinados a microempresas y pequeñas empresas. En la Figura 1-2 se muestran los montos de créditos a pequeñas empresas que han sido otorgados por mes en las diversas entidades financieras, pudiendo destacarse un punto en el mes de setiembre del 2017 en el que este monto llegó a sobrepasar los 26 000 millones de soles. Esta Figura muestra además que desde el mes de setiembre de 2015 los montos otorgados han ido en un aumento considerable y hasta marzo de 2018 se han concedido alrededor de 26 000 millones de soles, y esta cifra va en aumento según se puede observar en el comportamiento de la gráfica.

En la Figura 1-3 se muestran los montos de créditos a Micro y Pequeñas Empresas (MyPEs) que han sido otorgados por mes en las diversas entidades financieras, allí se observa que los

montos han ido creciendo desde febrero de 2016 con aproximadamente 32 000 millones de soles hasta sobrepasar los 38 000 millones en febrero de 2019.

Figura 1-3. Montos otorgados por microcréditos a MyPE hasta febrero de 2019

Fuente: Superintendencia de Banca Seguros y AFP



Los microcréditos poseen un importante índice de concentración de mercado, quiere decir que este tipo de producto tiene un alto poder, y también representa mayores ganancias para las entidades financieras que la ofrecen.

Según la Figura 1-3 sobre:

Los índices de concentración de Herfindhal IHH, el sistema financiero peruano muestra elevada concentración, tanto a nivel agregado, como de cada tipo de crédito. […] Las CMAC también muestran alta concentración en microcrédito, así como en crédito corporativo, lo que refleja su especialización en estos productos y sus clientes. (León & Jopen, 2011, p. 230)

Los prestamos para el consumo, micro y pequeña empresa forman parte de los microcréditos, el cual representa el 35% de los prestamos totales emitidos por las entidades financieras a febrero de 2019 según lo mostrado por la Figura 1-4.

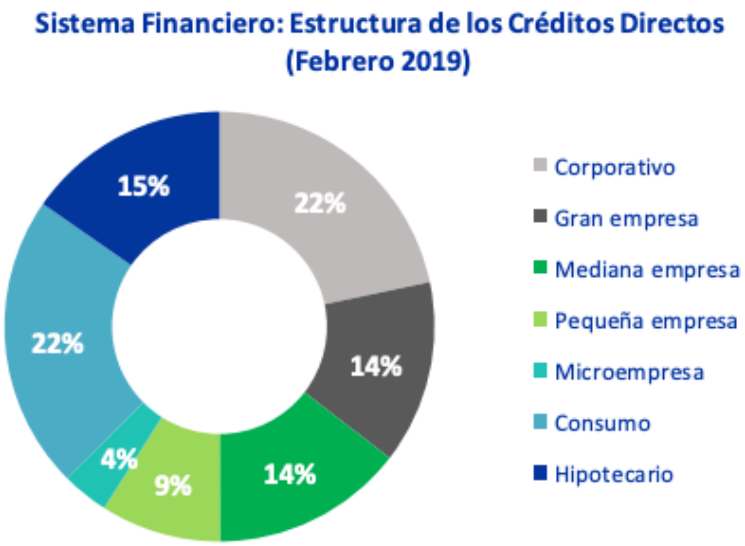


Figura 1-4. Sistema Financiero: Estructura de los Créditos Directos (Febrero 2019)

Fuente: Superintendencia de Banca Seguros y AFP

La deserción de clientes micro-créditos en la banca peruana va desde los 7% hasta los 15% según estudio realizado por Arellano Marketing en Noviembre del 2018, actualmente los bancos tienen una alta tasa de clientes no fidelizados (propensos a desertar), solo por detrás de las empresas de Telefonía móvil e Instituciones públicas, como se puede observar en la Figura 1-5.

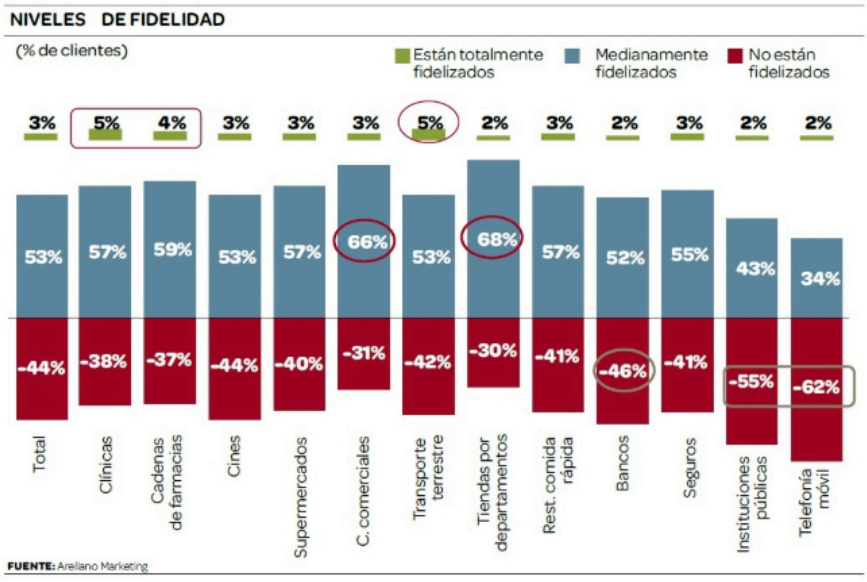


Figura 1-5. Nivel de fidelización en los diferentes sectores (Noviembre 2018)

Fuente: Arellano Marketing

El proceso de predicción de deserción de clientes en la banca facilitaría el trabajo del área de marketing de las entidades financieras, debido a que tendrían que enfocarse principalmente en los clientes que tienen más riesgo de deserción. Automatizar este proceso trae como consecuencia el ahorro de tiempo hombre del área de marketing y demás relacionados con el tema de deserción de clientes. Ahorro de costos debido a que al hacerlo de manera manual no se identifica con certeza el objetivo de productos especiales que tratan de fidelizar al cliente.

1.4 Objetivos

1.4.1 Objetivo principal

Desarrollar un sistema inteligente basado en un modelo hibrido para mejorar la predicción de deserción de clientes microcréditos en bancos de la ciudad de Lima.

1.4.2 Objetivos específicos

- Identificar los atributos a utilizar para el proceso de la predicción de deserción de clientes bancarios.
- Revisar y evaluar las técnicas de predicción de deserción de clientes bancarios.
- Identificar el modelo más adecuado para la predicción de deserción de clientes en la banca.
- Desarrollar el modulo para el modelo identificado
- Validar el modelo desarrollado y obtener una mejor precision

2 Capítulo 2. Revisión de métodos para la predicción de deserción de clientes bancarios

En este capítulo revisamos diversos métodos para la predicción de la deserción de clientes microcréditos en bancos.

En el presente capítulo se consideran los lineamientos utilizados por Kitchenham *et al.* (2007), las cuales se adaptaron para determinar los siguientes tres pasos.

- Planificación: en esta fase se elaboran las preguntas y se define el protocolo de investigación.
- Desarrollo: en esta fase se seleccionan los principales estudios de acuerdo a los criterios de inclusión y exclusión.
- Resultados: en esta fase se presentan las estadísticas y el análisis realizado a los estudios previamente seleccionados.

2.1 Planificación

Con el fin de conocer los métodos más adecuados para la predicción de la deserción de clientes bancarios no planteamos las siguientes preguntas.

- Q1: ¿Cuál es la técnica más adecuada para el balanceo de grandes cantidades de datos desproporcionales?
- Q2: ¿Qué factores son los más influyentes en los clientes para que los motive en la toma de decisión de dejar el servicio bancario?
- Q3: ¿Qué técnicas se han desarrollado para la predicción de deserción de clientes en la banca?
- Q4: ¿Cuál es la técnica más adecuada para abordar el problema de la predicción de deserción de clientes en la banca?

Las siguientes bases de datos fueron usadas principalmente para definir el protocolo de búsqueda: ACM, IEEE, Springer, Science Direct. La investigación cubre el periodo de enero del 2008 hasta julio 2018.

En la investigación se utilizó los siguientes criterios de búsqueda TITLE-ABS-KEY (“Churn Forecasting Banking”) o "TITLEABS-KEY" (“Desertion Prediction Financial Services”) o TITLEABS-KEY (“Churn Prediction Credit Card”) que fueron aplicados en filtros de título, resumen y palabras clave.

Posteriormente, se consideraron los siguientes criterios de inclusión:

- Comparación de técnicas que ayuden a la solución del problema de deserción de clientes bancarios.
- Factores propuestos que influencien la deserción de clientes.
- Estudios relacionados con deserción de clientes en el sistema financiero.

Y los siguientes criterios de exclusión:

- Modelos estadísticos y de simulación.
- Deserción de clientes en sectores de telecomunicaciones, seguros o empresas de *retail.*
- Journals que presenten factor de impacto

2.2 Desarrollo

La metodología de búsqueda se realizó en los siguientes bancos de papers y journals: Springer, Science Direct, IEEE y ACM, en los cuales se aplicaron los descriptores (*Keywords*) mencionados anteriormente, de los cuales se obtuvieron 554 artículos; la cantidad total de artículos no eran especializados en el problema, por lo que se aplicó un filtro por los criterios de inclusión, lo cual redujo la cantidad de resultados a 115 artículos; posteriormente se aplicó un segundo filtro mediante los criterios de exclusión. Al finalizar este proceso de búsqueda se obtuvieron 25 papers especializados en el problema de deserción de clientes bancarios.

El proceso de selección de artículos se presenta en la Figura 2-1.

El resultado de la aplicación de la metodología de búsqueda dio como resultado un total de 35 papers, de los cuales para este estudio se tomaron solo 25 que se muestran en la Fuente: elaboración propia

a 2-1, debido a que en ellos se presentan técnicas novedosas, evaluación de factores influyentes, revisión de la literatura y comparación de técnicas.

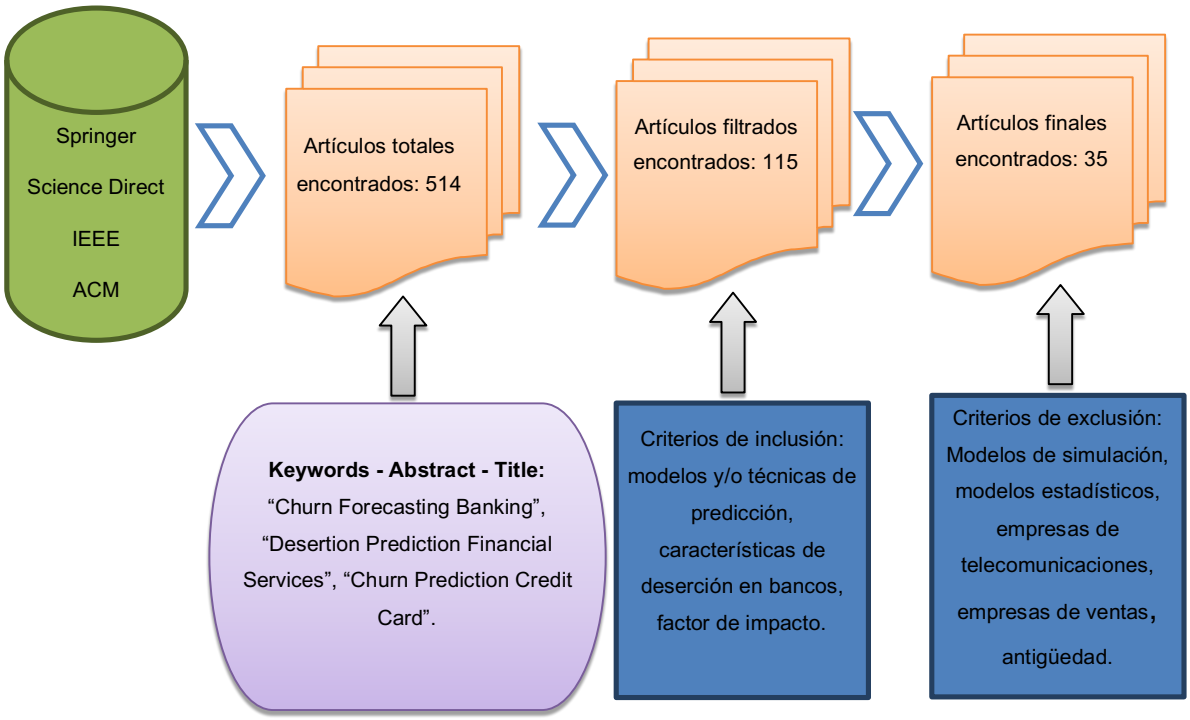


Figura 2-1. Resultados de la metodología de búsqueda

Fuente: elaboración propia

Tabla 2-1. *Relación de papers estudiados*

Autor	Año	Título	Journal
Abbas Keramati, Hajar Ghaneei and Seyed Mohammad Mirmohammadi	2016	Developing a prediction model for customer churn from electronic banking services using data mining	Financial Innovation
Sebastián Maldonado	2015	Churn prediction via support vector classification: An empirical comparison	Intelligent DataAnalysis
Bing Zhu and Jin Xiao	2014	A Balanced Transfer Learning Model for Customer Churn Prediction	Proceedings of the Eighth International Conference on Management Science and Engineering Management
Chiun-Sin Lin, Gwo-Hshiung Tzeng, Yang-Chieh Chin	2011	Combined rough set theory and flow network graph to predict customer churn in credit card accounts	Expert Systems with Applications
Dries F. Benoit, Dirk Van den Poel	2012	Improving customer retention in financial services using kinship network information	Expert Systems with Applications
Guangli Nie, Wei Rowe, Lingling Zhang, Yingjie Tian, Yong Shi	2011	Credit card churn forecasting by logistic regression and decision tree	Expert Systems with Applications
G. Ganesh Sundarkumar, Vadlamani Ravi	2015	A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance	Engineering Applications of Artificial Intelligence

Jin Xiao, Xiaoji Jiang, Changzheng He, Geer Teng	2016	Churn Prediction in Customer Relationship management via GMDH	IEEE
Leilei Tang, Lyn Thomas, Mary Fletcher, Jiazhu Pan, Andrew Marshall	2014	Assessing the impact of derived behavior information on customer attrition in the financial service industry	European Journal of Operational Research
M.A.H Farquad, V. Ravi, S. Bapi Raju	2012	Analytical CRM in banking and finance using SVM	Electronic Customer Relationship management
M.A.H. Farquad, Vadlamani Ravi, S. Bapi Raju	2014	Churn prediction using comprehensible support vector machine: An analytical CRM application	Applied Soft Computing
M. C. López-Díaz, M. López-Díaz S. Martínez-Fernández	2016	A stochastic comparison of customer classifiers with an application to customer attrition in commercial banking	Scandinavian Actuarial Journal
Ning Wang, Dong-xiao Niu	2009	Credit Card Customer Churn Prediction Based on the RST and LS-SVM	IEEE
Özden Gür Ali, Umut Arıtürk	2014	Dynamic churn prediction framework with more effective use of rare event data: The case of private banking	Expert Systems with Applications
Saran Kumar A. Chandrakala	2016	A Survey on Customer Churn Prediction using Machine Learning Techniques	International Journal of Computer Applications

Sebastián Maldonado	2015	Churn prediction via support vector classification: An empirical comparison	Intelligent Data Analysis
Xi Zhao, Yong Shi, Jongwon Lee, Heung Kee Kim, Heeseok Lee	2014	Customer Churn Prediction Based on Feature Clustering and Nonparallel Support Vector Machine	International Journal of Information Technology & Decision Making
Yaya Xie, Xiu Li, E.W.T. Ngai, Weiyun Ying	2009	Customer churn prediction using improved balanced random forests	Expert Systems with Applications
Yaya Xie, Xiu Li	2008	Churn Prediction with Linear Discriminant Boosting Algorithm	IEEE
Zhao Jing, Dang Xing-hua	2008	Bank Customer Churn Prediction Based on Support Vector Machine - Taking a Commercial Bank's VIP Customer Churn as the Example	IEEE
Renzo Barrueta, Jean Paul Castillo, Jimmy Armas	2018	Predictive model to determine customer desertion in Peruvian banking entities	IEEE
T. Vafeiadisa, K. I. Diamantaras, & G. Sarigiannidis	2015	A comparison of machine learning techniques for customer churn prediction	Simulation Modelling Practice and Theory
A. De Caignya, K. Coussementa, & Koen W	2018	A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees	European Journal of Operational Research
K. Ravi, V. Ravi, & P. Sree Rama Krishna	2017	Fuzzy Formal Concept Analysis based Opinion Mining for CRM in Financial Services	Applied Soft Computing

F. Shirazia, & M. Mohammadi	2018	A big data analytics model for customer churn prediction in the retiree segment	International Journal of Information Management

Fuente: elaboración propia

Con respecto a la investigación realizada por el motor de búsqueda Scopus se muestran los resultados en el Figura 2-2:

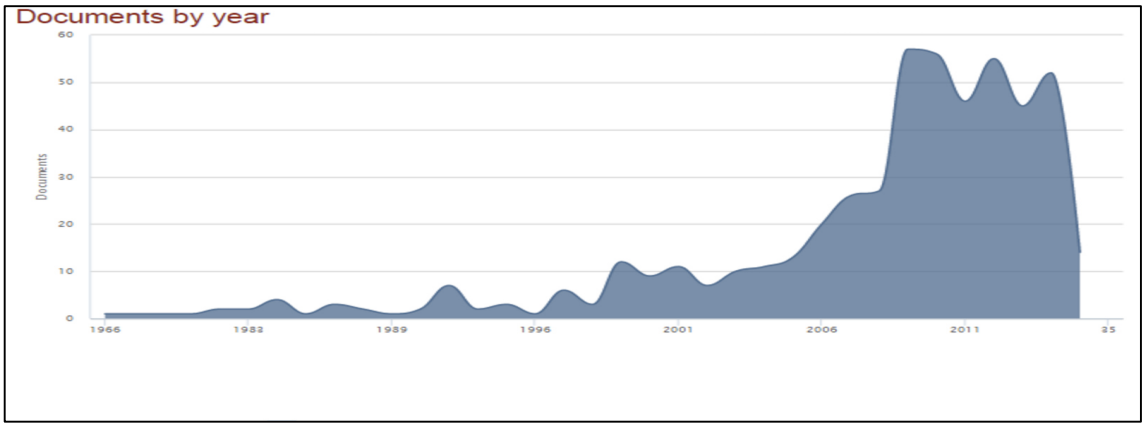


Figura 2-2. Cantidad de estudios publicados por año (Scopus)

Fuente: elaboración propia

En este gráfico se puede observar la cantidad de papers publicados por año; el primer estudio fue realizado en el año 1966, la cantidad de estudios relacionados con el tema de la deserción de clientes bancarios ha ido incrementando posteriormente, más aún en los últimos 9 años (2009 al 2018) en donde la cantidad de publicaciones llegó a ser de 57 como máximo en el año 2009, lo cual demuestra el interés por la deserción de clientes en la banca.

2.3 Resultados

En esta sección se le da respuesta a cada una de las preguntas de investigación propuestas en la sección 2.1.

- **Q1: ¿Cuál es la técnica más adecuada para el balanceo de grandes cantidades de datos desproporcionales?**

En el problema de deserción de clientes bancarios se cuenta con 2 clases (desertor, no desertor) altamente desbalanceadas, aproximadamente se encuentra en la proporción de 1 a 10

entre la clase desertora y la no desertora correspondientemente. Para que los métodos de predicción puedan procesar con mayor eficiencia las muestras de datos se aplican diversas técnicas de balanceo, las cuales se muestran en la Tabla 2-2.

Tabla 2-2. *Relación métodos para balanceo de data*

Técnica	Descripción	Autor
Balanced Random Forest	Artificialmente hace las clases iguales por un exceso de muestreo (over sampling) de la clase minoritaria en el aprendizaje extremadamente desequilibrada	(Xie, Li, Ngai, & Ying, 2009)
Random under-sampling	Under-sampling es una técnica en la que algunas de las muestras que pertenecen a la clase de la mayoría se eliminan al azar y se combinan con las muestras de clase minoría.	[Farquad, Ravi, y Bapi (2014)- el mejor resultado], [Nie, Rowe, Zhang, Tian y Shi (2011)- el mejor resultado], (Keramati, Ghaneei, & Mohammad, 2016; Maldonado, 2015; Zhu, Xiao, & He, 2014)
Random over-sampling	Over-sampling es una técnica en la que las muestras que pertenecen a la clase minoritaria se replican varias veces y se combinaron con las muestras de la clase mayoritaria.	(Farquad *et. al*, 2014; Xiao, Jiang, He, & Teng, 2016)
Synthetic minority over-sampling technique (SMOTE)	SMOTE es un enfoque en el que la clase minoritaria es sobre muestreada mediante la creación de muestras sintéticas (o artificiales), en lugar de por sobre muestreo con reemplazo. La clase minoritaria es sobre muestreada sacando cada muestra y la introducción de muestras sintéticas a lo largo de los segmentos de línea que unen cualquier / todos los vecinos más cercanos de la clase minoritaria k.	(Farquad *et al*., 2014; Maldonado, 2015; Zhu *et al*., 2014; Maldonado, 2015)
KReverse Nearest Neghbor	Consiste en la eliminación de datos con similares características entre un conjunto de puntos conglomerados (vecinos).	(Sundarkumar & Ravi, 2015)

One Class Support Vector Machine	OCSVM es diferente de la SVM en el sentido de que los datos de entrenamiento pertenecen a una sola clase. Se construye una frontera que separa la clase desde el resto de las características.	(Sundarkumar & Ravi, 2015)
ALBA	ALBA hace uso de los vectores de soporte para generar muestras adicionales cerca de ellos.	(Farquad, Ravi, & Bapi, 2012)

Fuente: elaboración propia

De la revisión de la literatura podemos observar que los autores que consideran técnicas de balanceo presentan una mejora de hasta 40% en sus resultados. Existen diversos métodos que se emplean para el tratamiento de las clases desbalanceadas, entre las más usadas se encuentran el Random Under Sampling y el SMOTE.

- **Q2: ¿Qué factores son los más influyentes en los clientes para que los motive en la toma de decisión de dejar el servicio bancario?**

Las características son los factores de entrada más importantes que se presentan en el modelo, debido a que la información con respecto al problema será ingresada mediante ellas, por eso se debe tener en consideración cuáles son decisivas.

De acuerdo con la literatura revisada (estudio de papers seleccionados) se tiene la Tabla 2-3 con los factores que se consideran para cada cliente en el problema de la deserción de clientes bancarios.

Tabla 2-3. *Relación de factores tomadas en cuenta en los papers revisados*

Id	Factor	Descripción	Referencia
F1	Edad	Edad del cliente	(Xie *et al*., 2009; Sin, Hshiung, & Chieh, 2011; Farquad *et al.,* 2014; Benoit & Van den, 2012; Tang, Thomas, Fletcher, Pan, & Marshall, 2014; Nie *et al.,* 2011; Wang & Niu, 2009; Xie & Li, 2008; Ali *et al*. 2014; Zhao, Shi, Lee, Kee, & Lee, 2014; Keramati, Ghaneei, & Mohammad, 2016; Farquad *et al.*, 2012),[(López, López, & Martínez, 2016)]
F2	Género	Género del cliente	(Sin *et al.,* 2011; Farquad et al., 2014; Benoit & Van den, 2012; Tang *et al.,* 2014; Nie *et al.,* 2011; Xie & Li, 2008; Ali *et al*. 2014; Zhao *et al*., 2014; Keramati *et al*., 2016; Zhu *et al.,* 2014; Farquad *et al.,* 2012; Xiao et al., 2016)
F3	Educación	Grado de educación del cliente	(Xie *et al*., 2009; Sin *et al.,* 2011; Farquad *et al.,* 2014; Nie *et al*., 2011; Xie & Li, 2008; Ali *et al*. 2014; Zhao *et al*., 2014; Keramati *et al.,* 2016; Zhu *et al.,* 2014; Farquad *et al.*, 2012; Xiao *et al.,* 2016)
F4	Ingresos disponibles anuales	Cantidad de dinero disponible en un año	(Xie *et al*., 2009; Sin *et al.,* 2011; Farquad *et al*., 2014; Benoit & Van den, 2012; Sundarkumar & Ravi, 2015; Wang & Niu, 2009; Xie & Li, 2008; Zhao *et al.,* 2014; Farquad *et al.*, 2012; Xiao *et al.,* 2016)
F5	Tipo de empleo	Sector al que pertenece el cliente	(Xie *et al*., 2009; Tang *et al.,* 2014; Zhao *et al.,* 2014; Xiao *et al.,* 2016)
F6	Estado civil	Condición del cliente según el registro civil	(Xie *et al*., 2009; Sin *et al.,* 2011; Farquad *et al*., 2014; Nie *et al.,* 2011; Gür & Arıtürk, 2014; Zhao *et al.,* 2014; Farquad *et al.*, 2012)
F7	Número de hijos	Número de hijos que tiene el cliente	(Xie *et al*., 2009)
F8	Estado financiero	Informe del estado de cuenta	(Tang *et al*., 2014; López *et al*., 2016)
F9	Estado social	tipo de persona natural o jurídica	(Benoit & Van den, 2012; Xie & Li, 2008; Benoit & Van den, 2012; Keramati *et al*., 2016; Zhu *et al*., 2014)

F10	Grado de servicio	Grado satisfacción del cliente con el banco	(Xie *et al*., 2009; Nie *et al.,* 2011)
F11	Tipo de cuenta	Tipo de cuenta proporcionada por el banco	(Xie *et al*., 2009)
F12	Tipo de garantía	posesiones, bienes inmuebles	(Xie *et al*., 2009)
F13	Periodos de posesión de tarjeta de crédito	Periodos de posesión de tarjeta de crédito	(Sin *et al.,* 2011; Nie *et al.,* 2011; Jing & Xing, 2008; Wang & Niu, 2009; Benoit & Van den, 2012; Keramati *et al*., 2016)
F14	Tiempo desde el pedido del préstamo	Periodo transcurrido desde que el banco brindó la tarjeta	(Xie *et al*., 2009; Nie *et al.,* 2011; Benoit & Van den, 2012; Xiao *et al*., 2016)
F15	Monto del préstamo	Cantidad de dinero que le proporcionó el banco	(Xie *et al*., 2009; Sin *et al.,* 2011; Farquad *et al.,* 2014; Benoit & Van den, 2012; Nie *et al.,* 2011; Sundarkumar & Ravi, 2015; Wang & Niu, 2009; Xie & Li, 2008)
F16	Estado de la cuenta	Si la cuenta se encuentra activa, inactiva, bloqueada, etc.	(Xie *et al*., 2009; Benoit & Van den, 2012; Nie *et al.,* 2011, Jing & Xing, 2008; Xie & Li, 2008; López *et al.*, 2016)
F17	Estado de crédito	Estado en el que se encuentra el crédito (cancelado o en deuda)	(Xie *et al*., 2009; Benoit & Van den, 2012; Nie *et al.,* 2011, Jing & Xing, 2008; Wang & Niu, 2009; Xie & Li, 2008; López *et al.*, 2016)
F18	El número de incumplimiento de normas	Cantidad de veces que el cliente incumplió el reglamento	(Xie *et al*., 2009; Nie *et al.,* 2011)

F19	Transferencia de débito automático	Transferencia de débito automático	(Sin *et al.,* 2011; Nie *et al.,* 2011; Sundarkumar & Ravi, 2015; Keramati *et al*., 2016)
F20	Amortización anuales	Cantidades de amortización anuales	(Sin *et al.,* 2011; Nie *et al.,* 2011)
F21	Tiempo de sin haber transacciones	Tiempo en el cual no se realizaron transacciones	(Sin *et al.,* 2011; Nie *et al.,* 2011; Wang & Niu, 2009; Benoit & Van den, 2012; Maldonado, 2015)
F22	Promedio anual de consumo	Promedio de consumo del cliente durante un año	(Sin *et al.,* 2011; Wang & Niu, 2009; Benoit & Van den, 2012; Xiao *et al*., 2016; López *et al.,*2016)
F23	Cantidad de veces de compras anuales	Número de compras que realizó en un año	(Sin *et al.,* 2011; Benoit & Van den, 2012; Wang & Niu, 2009; Xie & Li, 2008; Gür & Arıtürk, 2014; Zhu *et al*., 2014)
F24	Consumo en los últimos tiempos	Monto de consumos recientes	(Sin *et al.,* 2011, Farquad *et al.,* 2014; Sundarkumar & Ravi, 2015; Wang & Niu, 2009; Zhu *et al*., 2014; Xiao *et al*., 2016)
F25	Número de tarjetas de crédito	Número de tarjetas de crédito que posee el cliente	(Farquad *et al.,* 2014; Nie *et al.,* 2011; Sundarkumar & Ravi, 2015; Maldonado, 2015)
F26	Transacciones web	Número de transacciones web	(Farquad *et al.,* 2014; Nie *et al.,* 2011; Sundarkumar & Ravi, 2015; Keramati *et al*., 2016); Maldonado, 2015; Farquad *et al*. 2012 Xiao *et al*., 2016; Maldonado, 2015)
F27	Transacciones vía móviles	Número de transacciones vía móviles	(Keramati *et al*., 2016)
F28	Fondos de inversión	Si el cliente posee fondos de inversión	(Xie & Li, 2008; Maldonado, 2015)

F29	Pago mensual	Pago mensual del crédito	(Tang *et al.,* 2014; Zhu *et al.,* 2014; Farquad *et al.,* 2012)
F30	Quejas de cliente	Cantidad de veces que el cliente se queja	(Nie *et al.,* 2011)
F31	VIP	Si el cliente es VIP	(Nie *et al.,* 2011)
F32	Monto usado	Proporción de monto usado con respecto al total	(Zhu *et al.,* 2014; López *et al.,* 2016)
F34	Atraso en cuotas	Si se atrasó en el último mes	(Xiao *et al*., 2016)
F35	Número de retiros	Número de veces que realizó un retiro	(Xiao *et al*., 2016), (López *et al.,* 2016)
F36	Compras con tarjetas de crédito	Número de compras realizada con tarjetas de crédito	(López *et al.,* 2016)
F37	Compras con tarjetas de débito	Número de compras realizada con tarjetas de débito	(López *et al.,* 2016)

Fuente: elaboración propia

Las características fueron obtenidas por lo general de una base de datos (Data warehouse), estas características han sido tomadas por cliente, es decir son las características para un solo cliente. En muchos casos se realiza la eliminación de variables por no ser de total influencia para la determinación del problema.

Del análisis del cuadro anteriormente mostrado se puede observar que muchos de los autores consideran sus propios factores para el tratamiento del problema y que solo coinciden en las variables de información básica del cliente.

Entre los factores más usados se tienen: Edad, Género, Educación, Ingresos, Estado civil, Periodo de posesión del crédito, Monto del crédito, Estado de la cuenta, Estado del crédito, Periodo de no transacciones y Cantidad de veces de consumo en un periodo.

- **Q3: ¿Qué técnicas se han desarrollado para la predicción de deserción de clientes en la banca?**

De acuerdo con la literatura revisada (estudio de papers seleccionados) se tiene la Tabla 2-4 con las técnicas usadas, además de una breve descripción de cómo se usó en el problema.

Tabla 2-4. *Relación métodos estudiados de los papers seleccionados*

Técnica	Descripción	Referencia
Improved Balanced Random Forests (IBRF)	Incorpora la técnica de muestreo Balanced Random Forest que tiene mayor eficiencia con gran cantidad de datos desbalanceados y la técnica de aprendizaje de coste-razonable, Weighted Random Forest que tiene un mayor efecto en la clasificación.	(Xie *et al.*, 2009)
Rough Set Theory (RST) and Flow Network Graph (FNG)	Se utiliza RST para extraer reglas relacionadas a la deserción de clientes, luego se usa el Flujo de Redes de Grafos para inferir reglas de decisión y variables.	(Sin *et al.,* 2011)
Comprehensible Support Vector Machine	Enfoque híbrido que combina la SVM-REF y NBT. SVM-REF (SVM-Eliminación de características recursivas) que es empleado para reducir la cantidad de características del conjunto de datos. Naive Bayes Tree es utilizado para la extracción de reglas luego de la aplicación de SVM-REF.	(Farquad *et al.*, 2014)
Random Forest	El autor realiza un agrupamiento de clientes, a los cuales agrupa y obtiene un conjunto de características sociales, estas características son utilizadas con la técnica Random Forest para obtener una clasificación de deserción.	(Benoit & Van den, 2012)
Aproximación Polinomial Ortogonal	Deriva la información indirectamente observable a partir de un conjunto de datos del panel único, que registra el valor de la historia financiera tenencias de política para cada cliente sobre una base mensual.	(Tang *et al.*, 2014)
Logistic Regression and Decision Tree	Se plantean los métodos de Logistic Regression y Decission Tree como técnicas para resolveré el problema de la deserción de clientes en bancos.	(Nie *et al.,* 2011)
Hybrid undersampling method using kReverse Nearest Neighbor and One Class SVM	Método híbrido para balancear la data utilizando kReverse Nearest Neighbor para la eliminación de valores similares que se encuentren cerca en cuanto a características, es decir uno representa a sus vecinos; y One Class SVM para obtener el vector de soporte que represente a la clase en estudio, de esta manera se eliminar valores inservibles.	(Sundarkumar & Ravi, 2015)

Support Vector Machine	Utilización de Support Vector Machine, mediante la utilización del Kernel de Cauchy.	(Jing & Xing, 2008)
Rough Set Theory and Leaf Squares – SVM	Utilización Rough Set Theory utilizado para refinar la data original mediante la discretización y eliminación de atributos; y la utilización de Leaf Squares – SVM, el cual transforma el problema de programación cuadrática de los SVM a un problema de ecuaciones lineales.	(Wang & Niu, 2009)
Linear Discriminant Boosting	Utilización de LDA para la creación de un patrón clasificador mediante las características de cada objeto.	(Xie & Li, 2008)
Dynamic intervals of time	Introduce la variable tiempo y precisa múltiples observaciones para un cliente, trabajando con las técnicas de Regresión Logística y Árboles de Decisión independientemente.	(Gür & Arıtürk, 2014)
MICAP-NPSVM	Híbrido entre las técnicas Coeficiente Máximo de Información, clusterización por afinidad y SVM no paralelos, entre sus bondades se tiene que no elimina registros con ruido sino que los agrupa entre los que presentan las mismas características.	(Zhao *et al.,* 2014)
SVM + NBTree	Durante la fase de aprendizaje se generan datos sintéticos, de los cuales se obtienen predicciones utilizando el modelo SVM desarrollado. Finalmente, durante la fase de generación de reglas, los algoritmos NBTree y DT.	(Farquad *et al.,* 2012)
GMDH	Modelo que puede seleccionar una solución de conjunto apropiada del grupo de clasificadores de forma adaptativa, determinar los pesos combinados entre clasificadores base seleccionados y completar el proceso de selección de conjunto automáticamente.	(Xiao *et al*., 2016)
Decision Tree + Forward Selection and Backward Elimination	Proporciona un modelo para la predicción de deserción basado en Árboles de decisión, previamente utiliza las técnicas de Forward Selection y Backward Elimination para conservar las variables más significativas al modelo.	(Keramati *et al.,* 2016)

Support Vector Data Descriptor	Variante de la técnica SVM, la cual aplica una variable R para la determinación de una esfera de clasificación, la cual separa las clases.	(Maldonado, 2015)
Decision Tree	Proporciona un modelo para la predicción de deserción basado en Árboles de decisión, tomando en cuenta 11 diferentes clases de clasificadores.	(López *et al.,* 2016)
SVM-SVDD	Estudió el modelo clásico SVM no lineal con núcleo gaussiano y SVDD no lineal con núcleo gaussiano.	(Maldonado, 2015)

Fuente: elaboración propia

De esta tabla se puede observar que los distintos autores utilizan diversas técnicas para la solución del problema de la predicción de deserción de clientes en servicios bancarios, además se encuentra en la literatura múltiples modelos para solucionar el problema.

En los estudios realizados se pueden observar técnicas híbridas en las cuales se unen las ventajas de dos métodos, usualmente cada uno de estos métodos se aplican en etapas diferentes de la solución del problema para abordar sub-problemas en particular.

En algunos casos se aplican técnicas derivadas de un origen común, por ejemplo muchos autores toman el concepto de las máquinas de soporte vectorial (SVM) para el desarrollo de sus estudios, como lo son: Leaf Squares – SVM, Comprehensible Support Vector Machine y One Class SVM.

- **Q4: ¿Cuál es la técnica más adecuada para abordar el problema de la predicción de deserción de clientes en la banca?**

Con la finalidad de responder a la pregunta, de la literatura se obtuvieron las métricas mostradas en la Tabla 2-5, estas métricas son utilizadas para la evaluación del desempeño de las técnicas de predicción encontradas en la literatura. Sin embargo la respuesta a la pregunta planteada se detallará en la Tabla 2-6.

Tabla 2-5. *Relación de métricas para el problema de la deserción de clientes en la banca*

Métrica	Descripción	Referencia
Lift curve (accuracy rate)	Para un umbral de probabilidad rotación dado, la curva de elevación traza la fracción de todos los abonados por encima del umbral en contra de la fracción de todos los desertores por encima del umbral.	(Xie *et al*., 2009; Jing & Xing, 2008; Xie & Li, 2008; Benoit *et al*., 2012; Zhao *et al.,* 2014; Keramati *et al.,* 2016; Maldonado, 2015; Zhu *et al*.,2014)
Top decile lift	Es el porcentaje del 10% de los clientes prevé que sea más probable para desertar que realmente deserto dividido por la tasa de rotación de referencia.	(Xie *et al*., 2009; Xie & Li, 2008; Gür & Arıtürk, 2014; Zhao *et al.,* 2014)
Sensitivity	La sensibilidad es la medida de la proporción de los verdaderos positivos (TP), que se identifican correctamente por clasificador.	(Farquad *et al*., 2014; Sundarkumar *et al*., 2015; Jing & Xing, 2008; Wang & Niu, 2009; Tang *et al.,* 2014; Keramati *et al*., 2016; Farquad *et al*., 2012; López *et al*. 2016)
Specificity	La especificidad es la medida de la proporción de los verdaderos negativos (TN), que se identifican correctamente por el clasificador.	(Farquad *et al.,* 2014; Sundarkumar *et al*., 2015; Jing & Xing, 2008; Wang & Niu, 2009; Farquad *et al*., 2012; Xiao *et al*., 2016; López *et al*. 2016)
Accuracy	La precisión es la medida de la proporción verdadera positivos y verdaderos negativos, que se identifican correctamente.	(Xie *et al*., 2009; Sin *et al*., 2011; Farquad *et al*., 2014; Nie *et al.,* 2011; Sundarkumar *et al.,* 2015; Jing & Xing, 2008; Wang & Niu, 2009; Gür & Arıtürk, 2014; Farquad *et al*., 2012; Keramati *et al.,* 2016; Zhu *et al*., 2014; Maldonado, 2015)
Misclassification cost	Es una función de coste para valorar los modelos construidos por los algoritmos de minería de datos que tienen los factores económicos en consideración.	(Nie *et al.,* 2011)

AUC	Es una métrica de la evaluación con frecuencia adoptada en la comunidad de la investigación para los problemas desequilibrados del aprendizaje	(Maldonado, 2015)

Fuente: elaboración propia

Se puede observar que entre las métricas más utilizadas se encuentran Sensitivity, Accuracy y Lift Curve, estas métricas son las comúnmente usadas, sin embargo, con el fin de determinar de una manera clara los resultados presentados en los papers se utilizará la métrica Accuracy (precisión).

En los papers estudiados se puede observar que todos los autores comparan su técnica propuesta con otras para demostrar su eficiencia, en la Fuente: elaboración propia

2-6 se presenta una comparación porcentual de la efectividad de los métodos según su precisión (accuracy).

Tabla 2-6. *Comparación de métodos presentados en los papers según exactitud*

Técnica	Dataset	Acierto	Referencia
Artificial Neural Network	20000	78.10%	(Xie *et al*., 2009)
	1504	54.70%	(Jing & Xing, 2008)
	1500	84.30%	(Wang & Niu, 2009)
	14814	72.30%	(Sundarkumar *et al*., 2015)
	1790	92.30%	(Zhu, *et al.,* 2014)
Decision Tree	20000	62.00%	(Xie *et al*., 2009)
	14814	60.60%	(Sundarkumar *et al*., 2015)
	1500	80.50%	(Wang & Niu, 2009)
	7204	75.00%	(Gür & Arıtürk, 2014)
	4283	96.70%	(Keramati *et al*., 2016)
Support Vector Machine	20000	87.20%	(Xie *et al*., 2009)
	14814	94.30%	(Farquad *et al.,* 2014)
	14814	60.40%	(Sundarkumar et. al, 2015)
	1504	59.70%	(Jing & Xing, 2008)
	1500	89.90%	(Wang & Niu, 2009)
	14000	87.60%	(Xiao *et al*., 2016)
	13812	50.00%	(Maldonado, 2015)
	1790	92.30%	(Zhu, *et al.,* 2014)
	52398	67,70%	(López *et al*., 2016)

Improved Balanced Random Forest	20000	93.20%	(Xie *et al*., 2009)
RST and flow Network Graph	21000	91.60%	(Sin e*t al.,* 2011)
SVM and Nayve Bayes Tree	14814	97.10%	(Farquad *et al*., 2014)
Nayve Bayes Tree	14814	96.50%	(Farquad *et al*., 2014)
	1504	55.40%	(Jing & Xing, 2008)
Logistic Regression	5456	84.70%	(Nie *et al.,* 2011)
	14814	60.80%	(Sundarkumar *et al.,* 2015)
	1500	58.90%	(Wang & Niu, 2009)
	7204	88.00%	(Gür & Arıtürk, 2014)
kRNN and One class Support Vector Machine	14814	78.10%	(Sundarkumar *et al.,* 2015)
Ridge Regression	1500	75.10%	(Wang & Niu, 2009)
Support Vector Data Descriptor	15816	74.50%	(Maldonado, 2015)

Fuente: elaboración propia

Mediante la comparación de los resultados obtenidos de los papers estudiados se concluye lo siguiente: las técnicas más usadas son las Máquinas de soporte vectorial (SVM), Árboles de decisión (DT), Redes neuronales artificiales (ANN) y Regresión logística (LR). Las técnicas que muestran mejores resultados son DT con un 96.7%, SVM y NBT presenta un 97.1%, IBRF con un 93.2% de exactitud, KRNN y OCSVM un 78.1%, SVM con una exactitud de 94.3%.

El porcentaje de acierto más alto alcanzado en la lista de papers revisados es el propuesto por (Farquad *et al*., 2014) quien utilizó SVM y NBT.

3 Capítulo 3. Revisión de KDD y técnicas de predicción

En el presente capítulo se realiza una descripción de la técnica general y las técnicas específicas utilizadas para la resolución del problema de deserción de clientes en la banca.

En la sección 3.1 se detalla la técnica seleccionada que servirá como guía de pasos para resolver el problema de minería de datos; posteriormente en la sección de Técnicas de predicción se describe cada una de las técnicas que se aplicarán posteriormente en la resolución del problema de deserción.

3.1 Knowledge Discovery in Databases

La técnica KDD es propuesta por Usama Fayyad, Gregory Piatetsky-Shapiro y Padhraic Smyth en el año 1996; está compuesta por cinco fases: selección, pre-procesamiento, transformación, minería de datos y evaluación e implantación, tal como se muestra en el Figura 3-1:

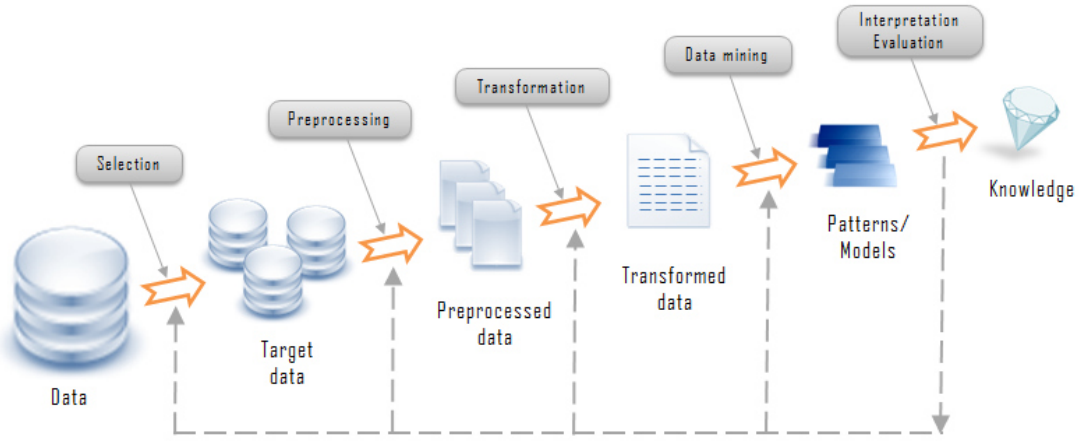


Figura 3-1. Knowledge Discovery in DataBases

Fuente: (WZL, s.f., párr. 3)

A continuación se hace una breve descripción de cada una de las fases de esta técnica.

- Selección: "Es la etapa en la que se determinan las fuentes de datos y tipos de información a utilizar […]. Los datos relevantes son extraídos de las distintas fuentes de datos" (WebMining Consultores, 2011, párr. 2) para las siguientes fases del KDD.

- Pre-procesamiento: esta etapa tiene como finalidad la preparación y la limpieza de datos obtenidos en la fase previa. Se utilizan diversas estrategias para manejar los datos faltantes, datos inconsistentes o fuera de rango. Como resultado final se obtiene un conjunto de datos adecuado para su posterior transformación (WebMining Consultores, 2011).
- Transformación: esta etapa consiste en el tratamiento preliminar de los datos, estos se transforman y generan nuevas variables a partir de las existentes (WebMining Consultores, 2011).
- Minería de datos: esta es la fase de modelamiento y aplicación de métodos inteligentes con la finalidad de extraer patrones previamente desconocidos y ocultos en los datos, nuevos, comprensibles y potencialmente útiles (WebMining Consultores, 2011).
- Interpretación y evaluación: en esta última fase se identifican los patrones obtenidos que son de interés, estos se evalúan basándose en algunas métricas para la obtención de resultados finales (WebMining Consultores, 2011).

3.2 Técnicas de predicción

A continuación se describirán cada una de las tres técnicas seleccionadas; estas técnicas se seleccionaron en base a la mayor exactitud que presenta cada una de las técnicas referenciadas en la literatura (Fuente: elaboración propia

). Xie *et al.* (2009) encontró que los Random Forest tienen un potencial grande para la resolución de problemas de deserción y en este paper se obtuvo una eficiencia de 93.2%. En el estudio de Farquad *et al*. (2014) y Wang y Niu (2009) se notó que las Máquinas de Soporte Vectorial ofrecen una muy buena solución para este problema, con un 97.1% y 89.9%, respectivamente. Sin embargo las Redes Neuronales Artificiales ofrecen un enfoque interesante para el tratamiento del problema tratado como se puede observar en el estudio de Wang y Niu (2009) en donde llegó a alcanzar un 84.3%.

3.2.1 Máquina de Soporte Vectorial (SVM)

Las máquinas de soporte vectorial son una técnica para la solución de problemas de clasificación o reconocimiento de patrones propuesto por Vapnik en 1992. Esta técnica es ideal para la clasificación de gran cantidad de datos complejos y con ruido.

La idea principal del SVM es la fijación de un hiperplano que separa datos de entrada en subgrupos, este debe estar ubicado de tal manera que se maximice la distancia entre él y los valores de entrada más cercanos, estos valores más cercanos son los llamados vectores de soporte como se puede observar en la Figura 3-2.

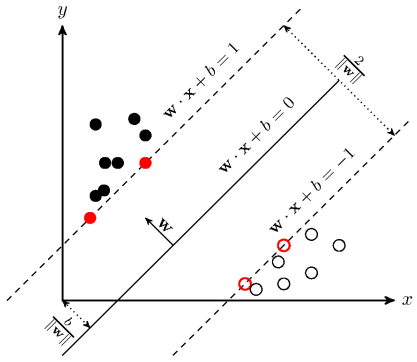


Figura 3-2. Máquina de soporte vectorial

Fuente: elaboración propia

Este problema de maximización se convierte en un problema de programación cuadrática, el cual se resuelve mediante los multiplicadores de Lagrange para posteriormente hallar la solución por su problema dual.

Existen tipos de problemas que se resuelven con las SVM

- Separador lineal: se aplica en el caso de tener los elementos del espacio inicial fácilmente separables sin necesidad de un procesamiento previo. El objetivo es el de maximizar la separación del hiperplano de los puntos de entrada más cercanos.
- Separador no lineal: se presentan dos problemas principales.
    - Datos separables con un margen máximo en un especio de características: en esta ocasión se convierte el espacio de entrada a un espacio de características mediante la aplicación de una función Kernel.
    - Margen blando: se presenta cuando no es posible encontrar una transformación de datos que permita separarlos linealmente, en este caso se permiten errores en la clasificación, pero estos son penalizados mediante la introducción de una constante C, determinada a priori, esta es una constante de holgura introducida al problema de programación cuadrática.

Vectores de soporte

Los vectores de soporte son los vectores conformados por los puntos a través de los cuales pasan las dos líneas paralelas al hiperplano. Estos vectores de soporte tienen la misma distancia cada uno al hiperplano y la distancia entre ellos es la mayor posible.

Función Kernel

La función Kernel permite la trasformación de los datos de un espacio inicial a otro de mayor dimensión llamado espacio de características como se puede observar en el Figura 3-3. En la aplicación por parte de las máquinas de soporte vectorial se utilizan para poder hallar un hiperplano en un espacio de una dimensión mayor.

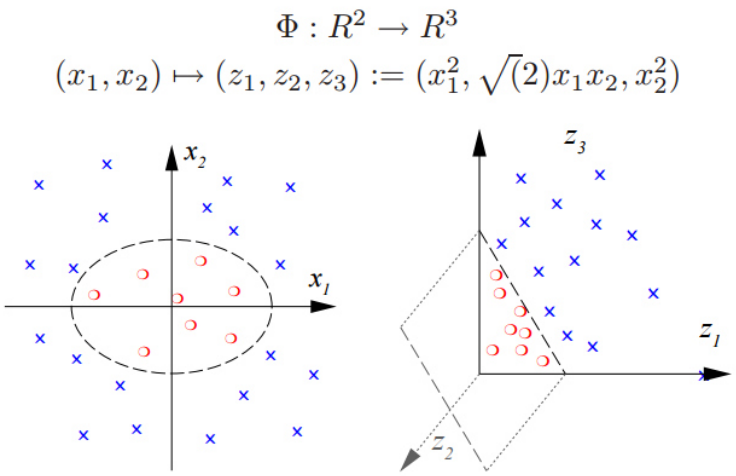


Figura 3-3. Función Kernel

Fuente: elaboración propia

Entre las funciones Kernel más usadas están:

- Kernel polinomial

$$K(x, y) = \left(\sum_{i=1}^{n} x_i y_i + 1 \right)^u$$

- Kernel RBF

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

- Kernel Cauchy

$$K(x, y) = \prod_{i=1}^{n} \frac{1}{1 + u(x_i - y_i)^2}$$

- Kernel Sigmoidea

$$k(x, y) = \tanh(\alpha x^T y + c)$$

- Kernel Laplace

$$k(x, y) = \exp\left(-\frac{\|x - y\|}{\sigma}\right)$$

Penalidad (Parámetro C)

El parámetro C indica qué tanto se quiere evadir el error de clasificación de cada ejemplo de entrenamiento en la optimización del SVM; para un gran valor de C, la optimización escogerá un hiperplano de menor margen si este logra hacer un mejor trabajo en separar correctamente todos los puntos de entrenamiento; por otro lado, un valor pequeño de C hará que el optimizador busque un hiperplano de separación de mayor margen incluso si este clasifica incorrectamente más puntos.

De la literatura revisada se obtuvo que el Kernel Cauchy con el parámetro u=0.3 y c=11 obtuvo el mejor resultado en comparación a los otros según el paper de Jing y Xing (2008), por otra parte en el paper de Wang y Niu (2009) se hizo una comparación similar que en el paper anteriormente citado y se encontró que la mejor función kernel es la RBF con el parámetro d=0.28 y c=10.

3.2.2 Random Forest

Los Random Forest (Bosques aleatorios) "son basados en árboles de decisión, clasificadores de instancias (registros de datos) representados como vectores de características. Los nodos prueban características, existe una rama para cada valor de la característica, las hojas especifican la categoría" (Arredondo, 2008, p. 4), las distintas clasificaciones se "pueden representar cualquier conjunción (AND) y disyunción (OR), además pueden representar

cualquier función de clasificación de vectores de características discretas" (Arredondo, 2008, p. 4).

Al igual que en los árboles de decisión (Figura 3-4) se pueden generar reglas a partir de los resultados obtenidos en las hojas (datos clasificados).

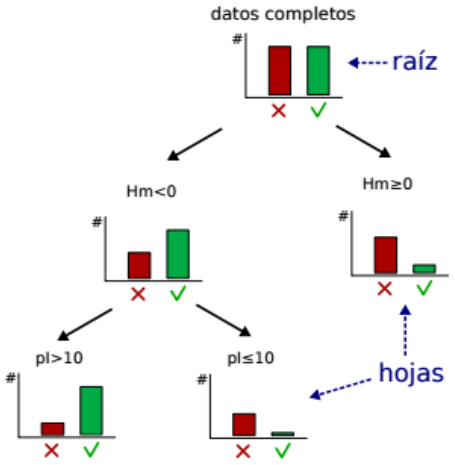


Figura 3-4. Árboles de decisión

Fuente: elaboración propia

La técnica de árboles aleatorios (RF) consiste en generar N árboles de decisión con M atributos del total de ellas. Estos N árboles de decisión son generados del total de ejemplos que se tienen (Figura 3-5), se divide el total entre los N árboles, de donde se tienen N subconjuntos para trabajar cada uno con un árbol con atributos distintos.

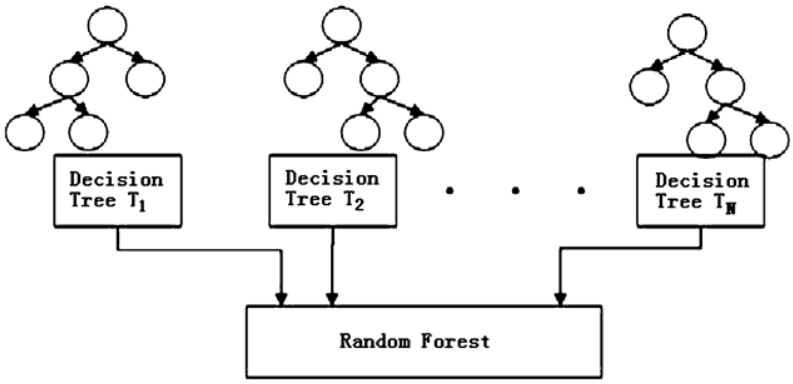


Figura 3-5. Random Forest

Fuente: elaboración propia

El clasificador del Random Forest se basa en un clasificador fuerte. Existen dos técnicas para crear un clasificador fuerte en base a clasificadores débiles. El Bagging y el Boosting.

Boosting: se tienen varios clasificadores débiles, los cuales son combinados finalmente para obtener un clasificador fuerte.

Bootstrap Aggregating (Bagging): entrena cada clasificador débil con un conjunto de datos diferente. Los conjuntos de datos se generan con el método de bootstrapping.

El bootstrap significa generar nuevos conjuntos de datos con el mismo tamaño usando un muestreo con reemplazamiento. En promedio una muestra bootstrap contiene 62,3 % de los muestras originales.

3.2.3 Redes Neuronales Artificiales (ANN) – Perceptrón Multicapa con Retro Propagación

El primer modelo de red neuronal fue propuesto en 1943 por McCulloch y Pitts en términos de un modelo computacional de actividad nerviosa. Este modelo era un modelo binario, donde cada neurona tenía un escalón o umbral prefijado, y sirvió de base para los modelos posteriores.

[…] Las características más importantes son:

- Auto-Organización y Adaptabilidad: utilizan algoritmos de aprendizaje adaptativo y auto-organización, por lo que ofrecen mejores posibilidades de procesado robusto y adaptativo.
- Procesado no Lineal: aumenta la capacidad de la red para aproximar funciones, clasificar patrones y aumenta su inmunidad frente al ruido.
- Procesado Paralelo: normalmente se usa un gran número de nodos de procesado, con alto nivel de interconectividad. (Marín, 2012, p. 5)

Una red neuronal consta de una capa de entrada, una capa oculta y una capa de salida (Figura 3-6).

La capa de entrada sirve para la presentación de datos a la red. En la capa oculta se encuentran como entradas todas las salidas de la capa de entrada, las cuales se ven afectadas

por un peso (intensidad). De igual manera en la capa de salida se obtienen las entradas de la capa oculta, esta última capa es la que da como salida el resultado de la red

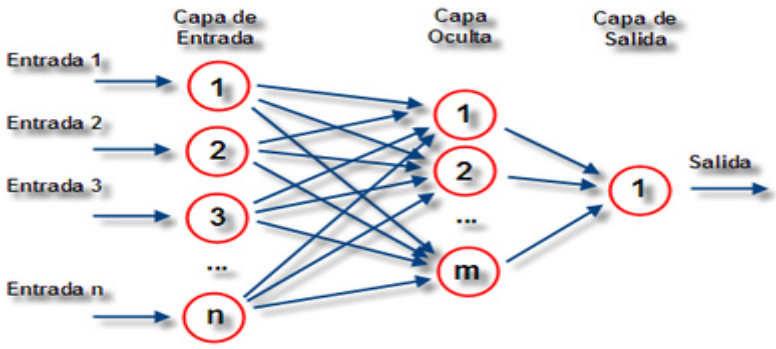


Figura 3-6. Redes Neuronales

Fuente: (Yepes, 2017, párr. 4)

Cada neurona posee un conjunto de entradas las cuales se ven afectadas por un peso asociado a cada entrada, estos pesos son modificados durante el proceso de aprendizaje de la red, estos pesos se van ajustando según la técnica seleccionada (Figura 3-7).

La neurona presenta una sumatoria de las entradas multiplicada por los pesos, a esta sumatoria se le conoce como función de propagación; el valor de esta función alimenta a la función de activación la cual produce la salida de la neurona. Esta función de activación produce diferentes resultados según la función que se escoja, entre las más conocidas están: Función Sigmoidea, Gaussiana, Sinusoidal, Escalón, etc.

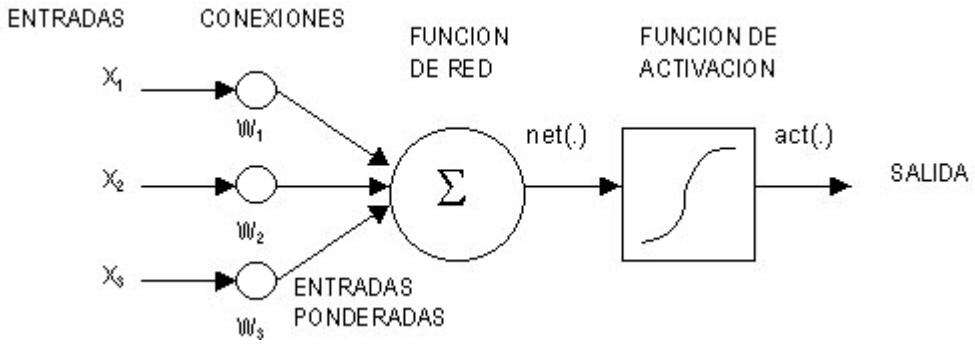


Figura 3-7. Neurona y sus componentes

Fuente: (Domínguez, s.f., párr. 1)

Perceptrón multicapa:

"El Perceptrón multicapa es una red neuronal artificial formada por múltiples capas, esto le permite resolver problemas que no son linealmente separables" (Xanadu Linux, 2017, párr. 15). Está compuesta por una capa de entrada, una de salida y una o más capas ocultas.

El aprendizaje de la red, vale decir la modificación de los pesos en las distintas iteraciones, se utiliza la retro propagación.

Retro propagación

Es un algoritmo de aprendizaje supervisado para el entrenamiento de redes multicapa. Posee dos fases para el aprendizaje:

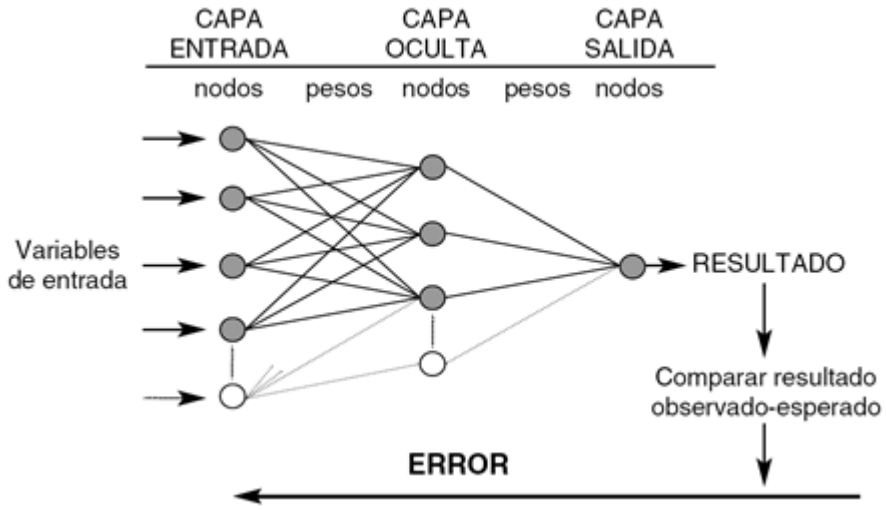


Figura 3-8. Perceptrón multicapa – Back Propagation

Fuente: (Trujillano *et al*., 2005, p. 15)

- Fase hacia adelante

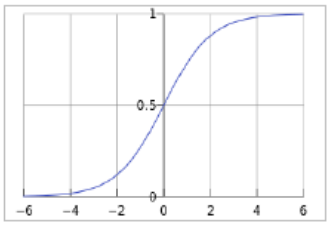
Los patrones de entrada son presentadas a la primera capa, la cual propaga el estímulo hacia las demás capas para obtener finalmente una salida de la red. Matemáticamente se tiene que cada una de las entradas netas que recibe una neurona es así:

$$u_j^p = \sum_{i=1}^{N} w_{ji}\, x_i^p + \theta_j$$

El valor de salida de la neurona se obtiene mediante la aplicación de una función, para el caso del perceptrón multicapa se aplica una de las dos siguientes funciones:

Función sigmoide

$$f(u_j) = \frac{1}{1+e^{-u_j}}$$



Función tangente hiperbólica

$$f(u_j) = \frac{e^{u_j} - e^{-u_j}}{e^{u_j} + e^{-u_j}}$$

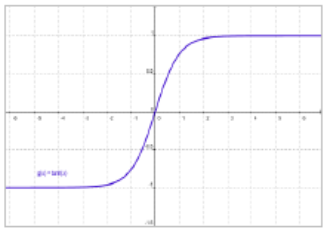


Figura 3-9. Funciones de salida

Fuente: elaboración propia

- Fase hacia atrás

El objetivo de esta fase es reducir al mínimo la diferencia entre la salida generada por la red y el patrón de salida (Figura 3-9). Esto se calcula mediante el error cuadrático medio, definido de la siguiente manera:

$$E_p = \frac{1}{2}\sum_{k=1}^{M}\left(t_k^p - y_k^p\right)^2 \Rightarrow E = \sum_{p=1}^{P} E_p$$

Para la modificación de los pesos (Aprendizaje) en esta fase se utiliza el Back Propagation la cual se basa en la Gradiente Descendiente. Estos pesos se modifican para disminuir el error cuadrático medio, para esto se cuenta con dos fórmulas que permiten la modificación de los pesos de la capa oculta; se tiene que su incremento está definido de la siguiente manera:

$$\Delta w_{ji} = \alpha \sum_{k}\left[(t_k - y_k)\, f'(u_k) w_{kj}\right] f'(u_j) x_i$$

Mientras que para el incremento de los pesos de las neuronas de la capa de salida se tiene:

$$\boldsymbol{\Delta w_{kj} = \alpha(t_k - y_k)\, f'(u_k) y_j}$$

Estos pesos se actualizan si es que el error que se encuentra es mayor a un error definido previamente y mientras que la cantidad de épocas no se haya superado.

4 Capítulo 4. Modelo propuesto para la predicción de deserción de clientes en la banca

En el presente capítulo se desarrolla el modelo para la deserción de clientes en banca siguiendo la metodología general "Knowledge Discovery in Databases" (KDD). La propuesta se basa en redes neuronales (RNA), máquinas de soporte vectorial (SVM) y árboles de decisión (DT); las cuales fueron descritas en el capítulo anterior. Además, los datos que se muestran en esta sección se utilizan para el entrenamiento del modelo.

4.1 Diseño de la investigación

El problema de predicción de deserción de clientes en la banca es un problema de minería de datos, donde se busca identificar patrones de comportamiento que definan cuando un cliente va a ser desertor o no. Para la solución de este problema se aplicará el KDD, el cual es un estándar de facto para los problemas de minería de datos; el KDD fija una serie de pasos que ayudarán en el proceso para encontrar un modelo que sea útil para identificar si un cliente será desertor o no en base a ciertas características.

El método a usar para desarrollar el modelo propuesto es Knowledge Discovery in Databases (KDD,) el cual es utilizado para problemas minería de datos (Predicción) como bien se señala en diversos trabajos (Miranda, 2006).

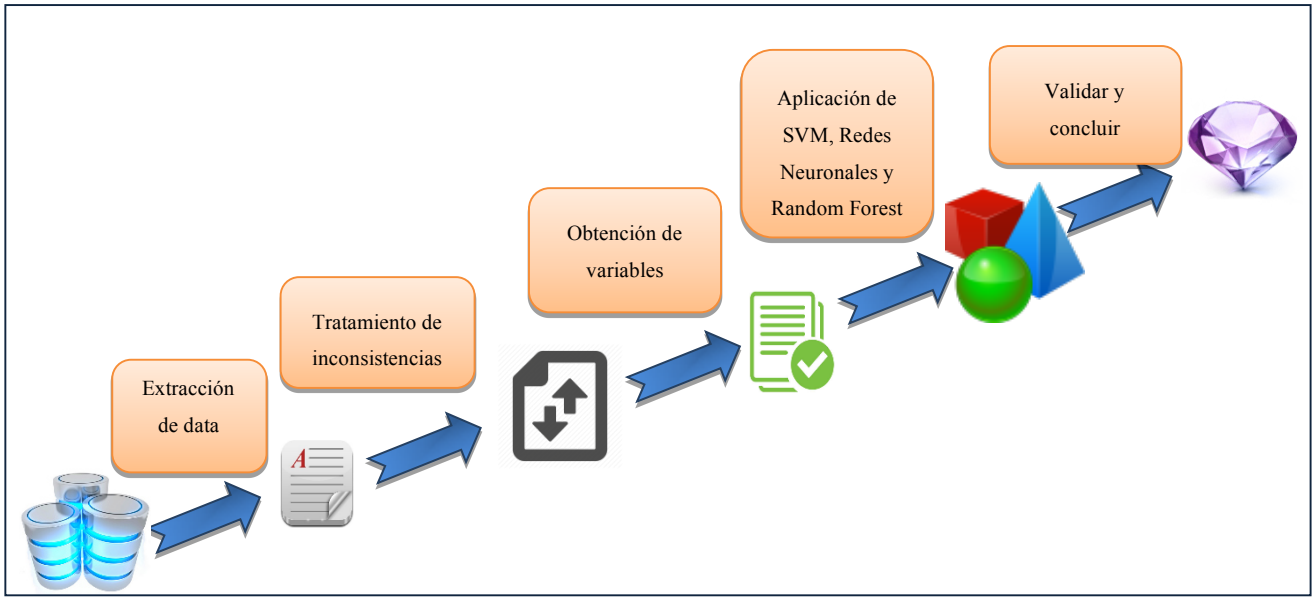


Figura 4-1. Esquema general de la solución de minería de datos mediante KDD

Fuente: elaboración propia

En esta sección se detalla cada uno de los cinco pasos mostrados en el Figura 4-1, los cuales se realizan para la aplicación del KDD en el problema de predicción de deserción de clientes en la banca. Como primer paso se tiene la extracción de la data (de un banco peruano) en el cual se definen los factores de los clientes. El siguiente paso es el tratamiento de inconsistencias de datos faltantes o con valores que no se encuentran en un rango definido. Posteriormente se realiza la conversión a variables numéricas y normalización en la cual se trabaja la data para obtener valores numéricos entre un rango determinado. Luego se elabora y afina el modelo de minería de datos, en este caso para las técnicas SVM, Random Forest y Redes Neuronales; en este capítulo solo se elabora el modelo. Se realiza la validación y las conclusiones, el detalle de este último paso se realiza en el Capítulo de Capítulo 6. Pruebas y validación y Capítulo de Capítulo 7. Conclusiones y trabajos futuros. Cada una de las 4 primeras fases se explica a continuación en las siguientes secciones.

4.2 Extracción de Data

En esta etapa se seleccionó el conjunto de datos utilizados como entrada para el modelo predictivo. La obtención de estos datos es fundamental para la elaboración de un modelo correcto, debido a que se debe contar con abundante cantidad de datos que contengan los factores que se mencionan en la Tabla 4-2. *Factores considerados para este estudio*de la sección 4.5.1 con la finalidad de poder realizar el entrenamiento y validación del modelo predictivo.

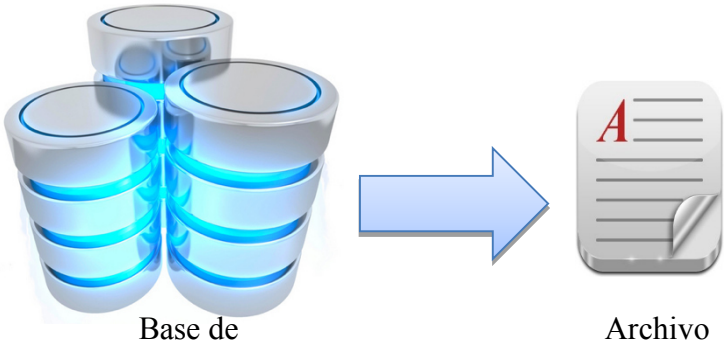


Figura 4-2. Selección de la data

Fuente: elaboración propia

Para el caso de estudio se analizó la data de los clientes de un banco peruano (Scotiabank), específicamente de clientes de micro-créditos que se encuentren domiciliados en la ciudad de lima, este dataset se encuentra conformado por 25000 registros de clientes.

El proceso mostrado en la Figura 4-2 representa la extracción de información de la base de datos a un archivo en formato CSV. Este último contiene datos de clientes bancarios para los cuales se tiene cada uno de los factores descritos en la sección 4.5.1, con cada uno de los nombres de ellas como cabecera del archivo de trabajo, como es mostrado en la Figura 4-3.

IDDESERTOR	EDAD	SEXO	EDUCACION	INGRESOS	RECLAMOS	POSECION	MONTO_PRESTAMO	NO_TRANSACCIONES	MONTO_CONSUMO1	MONTO_CONSUMO2	MONTO_CONSUMO3	MONTO_CONSUMO4	MONTO_CONSUMO5
23409	41	F	SUPERIOR	10700	0	15	9300	0	7254	7352	7088	7418	7294
23410	31	M	TECNICO	57008	0	33	28100	2	24300	23846	23777	23912	24167
23411	38	F	TECNICO	2500	0	9	3200	0	2211	1916	1351	1288	1579
23412	46	M	SUPERIOR	15102	0	16	24400	0	8004.96	7709	6030	4456	3250
23413	60	F	SECUNDARIA	2102	0	11	4400	0	2243	1459	1581.22	2217	2269

MONTO_CONSUMO6	TRANSACCIONES_MES1	TRANSACCIONES_MES2	TRANSACCIONES_MES3	TRANSACCIONES_MES4	TRANSACCIONES_MES5	TRANSACCIONES_MES6	DESERTOR
6850	4	4	4	5	4	3	NO
23795	3	3	1	3	5	3	NO
1951	4	5	3	5	1	3	NO
1150	5	4	3	4	1	1	SI
2147	8	7	8	5	5	4	NO

Figura 4-4. Datos de deserción de clientes en archivo CSV

Fuente: elaboración propia

4.3 Tratamiento de inconsistencias

En esta etapa se detectaron y trataron a las inconsistencias halladas en el conjunto de datos (archivo CSV obtenido como resultado de la fase anterior). Existen dos tipos de inconsistencias que se trabajaron: valores faltantes y valores fuera de rango.

Los valores faltantes son datos inexistentes en un registro asociado a una variable, mientras que los valores fuera de rango son los datos cuyo valor no se encuentra en el intervalo normal definido para esa variable; los rangos se definirán en la sección 4.5.1 Factores.

Tratamiento para los valores inconsistentes:

Si la cantidad de valores inconsistentes asociados a una variable es menor al 2% se eliminará el registro, es decir, el total de valores asociados a ese cliente. Por otro lado, si la cantidad de valores inconsistentes sobrepasa el 2% entonces estos valores se agregarán o reemplazarán con la media aritmética de todos los valores de esa variable (Miranda, 2006).

Tabla 4-1. *Criterios utilizados para resolver las inconsistencias*

Valor inconsistente	Acción
Menor a 2%	Eliminación del registro del cliente
Mayor a 2%	Reemplazo de los valores con el promedio

Fuente: elaboración propia

Se trataron las inconsistencias en el dataset de la siguiente manera:

- Se eliminaron 462 registros debido a que no presentaban fecha de nacimiento (factor que posteriormente se convertirá en edad).
- Se reemplazaron 3 158 registros con el promedio debido a que no presentaban grado de educación.

Por otro lado, como para este estudio solo se están considerando clientes micro-crédito se eliminaron aquellos que presenten créditos mayores a 100 000 soles, la cantidad de registros eliminados por este criterio fueron 358.

Al finalizar el tratamiento de inconsistencias se obtuvieron un total de 24 180 datos, los cuales serán utilizados para las siguientes etapas.

4.4 Obtención de nuevos factores

En esta etapa se trabajan los factores existentes para obtener nuevos que son determinantes para la precisión del modelo en el momento de predecir si un cliente es desertor o no.

Se plantea el uso de los siguientes factores:

- Cantidad de transacciones promedio (cantTxn)

Como resultado de promediar las variables cantTxn_1, cantTxn_2,…, cantTxn_6, las cuales son las cantidades de transacciones de los seis últimos meses.

- Monto de consumo promedio (montoTxn)

Como resultado de promediar las variables montoTxn_1, montoTxn_2,…, montoTxn_6, las cuales son los montos de transacciones de los seis últimos meses.

Debido a que los valores promedios reflejan el comportamiento más estable, a diferencia de los valores mensuales son más volátiles.

Para determinar qué tanto afectan estos nuevos factores al momento de predecir, se usó la correlación de Pearson, de las cuales solo "Cantidad de transacciones promedio" presenta una alta correlación; esta validación se describe en la sección 6.3.1. Finalmente el factor "Monto de consumo promedio" se descartó.

4.5 Aplicación de SVM, Redes Neuronales y Random Forest -Modelo de Minería de Datos

En esta sección se realizará una descripción de los factores considerados, la métrica seleccionada, el método de balanceo y los parámetros utilizados en las diferentes técnicas de minería de datos aplicados al modelo general que se plantea para la solución del problema de deserción de clientes en la banca.

4.5.1 Factores

Los factores son la principal entrada del sistema, representan la información que alimentará el modelo de predicción de deserción, además son de fundamental importancia para la resolución del problema.

En esta sección se describen los factores que fueron seleccionadas de acuerdo con la literatura revisada (estudio de papers seleccionados), en las cuales se consideraron los factores con mayor cantidad de referencias como se revisó anteriormente en la Fuente: elaboración propia

de la sección 2.3 Resultados.

Además se consideró el factor "Número de reclamos" a pesar de que no cuente con un alto número de referencias, debido a que aportó considerablemente a los resultados de Abbas, el cual fue el primero en considerarlo (Keramati *et al.,* 2016) presentando el mejor porcentaje de acierto de todos los papers revisados.

Por otro lado se tuvo en cuenta el factor propuesto en la sección 4.4 Obtención de nuevos factores, llamado "Cantidad de transacciones promedio".

En base a esto se tienen las variables mostradas en la Tabla 4-2 para la construcción del modelo de predicción de deserción de clientes en la banca.

Los factores seleccionados se clasificaron en tres grupos.

- Sociodemográficas: factores asociadas a los aspectos personales del cliente.
- Percepción del servicio: factores que miden el grado de satisfacción del cliente.
- Comportamiento bancario: factores que describen el comportamiento bancario de un cliente.

Los factores que se consideraran para este estudio agrupado por su clasificación son:

Factores	Descripción	Tipo	Valores Permitidos	Referencias
Sociodemográficas				
F1	Edad del cliente	DP	[+18]	(Xie *et al*., 2009; Sin *et al*., 2011; Farquad *et al*., 2014; Benoit & Van den, 2012; Tang *et al*., 2014; Nie *et al.,* 2011; Wang & Niu, 2009; Xie & Li, 2008; Ali *et al*., 2014; Zhao *et al*., 2014; Keramati *et al.,* 2016; Farquad *et al.,* 2012; López *et al*., 2016)
F2	Género del cliente	C	[M,F]	(Sin *et al*., 2011; Farquad *et al*., 2014; Benoit & Van den, 2012; Tang *et al*., 2014; Nie *et al.,* 2011; Xie & Li, 2008; Ali *et al*., 2014; Zhao *et al*., 2014; Keramati *et al.,* 2016; Farquad *et al.,* 2012; Xiao *et al*., 2016)
F3	Grado de educación del cliente	C	[Primaria, Secundaria, Técnica, Superior]	(Xie *et al*., 2009; Sin *et al*., 2011; Farquad *et al*., 2014; Nie *et al.,* 2011; Xie & Li, 2008; Ali *et al*., 2014; Zhao *et al.,* 2014; Keramati *et al.,* 2016; Zhu *et al*., 2014; Farquad *et al.,* 2012; Xiao *et al*., 2016)
F4	Ingresos disponibles Mensuales	CP	-	(Xie *et al*., 2009; Sin *et al*., 2011; Farquad *et al*., 2014; Benoit & Van den, 2012; Sundarkumar *et al*., 2015; Wang & Niu, 2009; Xie & Li, 2008; Zhao *et al.,* 2014; Farquad *et al*., 2012; Xiao *et al*., 2016)
Percepción de servicio				
F5	Número de reclamos	DP	-	(Keramati *et al.,* 2016)

Comportamiento bancario				
F6	Periodos (Meses) de posesión de tarjeta de crédito	DP	-	(Sin *et al*., 2011; Nie *et al.,* 2011; Jing & Xing, 2008; Wang & Niu, 2009; Benoit & Van den, 2012; Keramati *et al*., 2016; Zhu *et al*., 2014)
F7	Monto del crédito	CP	-	(Xie *et al*., 2009; Sin *et al.,* 2011; Farquad *et al.*, 2014; Benoit & Van den, 2012; Nie *et al.,* 2011; Sundarkumar *et al*., 2015; Wang & Niu, 2009; Xie & Li, 2008)
F8	Tiempo en meses de no realizar transacciones	DP	-	(Sin *et al.,* 2011; Nie *et al.,* 2011; Wang & Niu, 2009; Benoit & Van den, 2012; Maldonado, 2015)
F9… F14	Monto de las transacciones en el mes i (1 .. 6)	CP	-	(Sin *et al*., 2011; Farquad *et al.,* 2014; Sundarkumar *et al*., 2015; Wang & Niu, 2009; Zhu *et al.,* 2014; Xiao *et al*., 2016)
F15… F20	Cantidad de transacciones en el mes i (1 .. 6)	DP	-	(Sin *et al.,* 2011; Benoit & Van den, 2012; Wang & Niu, 2009; Xie & Li, 2008; Gür & Arıtürk, 2014; Zhu *et al*., 2014)
F21	Cantidad promedio de las transacciones mensuales	DP	-	Factor propuesto
Resultado				
F22	Cliente desertor o no en los próximos tres meses	C	[SI, NO]	

Tabla 4-2. *Factores considerados para este estudio*

Fuente: elaboración propia

Los factores mostrados en la Tabla 4-2 han sido seleccionados por ser los más referenciados en la Tabla 2-3 como se muestra en el análisis de dicha tabla, además de ser elegidos por estar incluidos en los experimentos realizados en los papers estudiados que presentan un mayor grado de acierto; también se agregó la variable "Número de reclamos" por los motivos presentados anteriormente. Los tipos C, CP y DP se refieren que son tipo categórica, continua positiva y discreta positiva, respectivamente.

Como se puede observar en la Tabla 4-2, se tienen variables categorias, las cuales se convertiran a valores enteros positivos como se muestra en la siguiente tabla :

Factores	Descripción	Valores Permitidos	Correspondencia de los valores
F2	Género del cliente	[M,F]	M=0 F=1
F3	Grado de educación del cliente	[Primaria, Secundaria, Técnica, Superior]	Primaria = 0 Secundaria = 1 Técnica = 2 Superior =3
Resultado			
F22	Cliente desertor o no en los próximos tres meses	[SI, NO]	SI = 1 NO = 0

Tabla 4-3. Conversion de valores de factores categoricas

Fuente: elaboración propia

Luego de tener todos los valores númericos de nuestros factores se realiza **la normalizacion de datos ,** en el rango de [0,1] denominda normalización basada en la unidad, usando la siguiente formula :

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

4.5.2 Balanceo

El balanceo de la data se describe como el proceso por el cual se equipará la cantidad de datos de dos o más conjuntos, en particular para el problema de la deserción se tiene que la data es altamente desbalanceada en términos de la proporción entre los desertores vs. los no desertores, teniendo aproximadamente un 7% Vs. 93%, respectivamente (Farquad *et al*., 2014).

Esta técnica fue escogida debido a que como se puede observar en la Q1 de la sección 2.3, presentar los mejores resultados en los papers estudiados de cuatro estudios que realizan balanceo dos de ellas presentan mejores resultados con esta técnica.

En el caso de estudio se cuenta con un total de 24 420 registros, de este total se cuenta con un porcentaje de desertores del 9.71% (2 372 clientes) versus el porcentaje de no desertores de 90.28% (22 048 clientes).

Tabla 4-3. *Método de balanceo utilizado en este estudio*

Técnica	Descripción
Random under-sampling 50%	Se toman los datos de la clase mayoritaria y se reduce al 50%

Fuente: elaboración propia

Esta técnica será aplicada al conjunto de datos de manera interna, es decir el programa recibirá toda la data y aplicará la técnica random under-sampling 50% para reducir la cantidad de datos de los clientes no desertores. Posteriormente el sistema tomará esta data, continuará con el procesamiento y aplicación de las técnicas.

4.5.3 Técnicas de predicción

Las técnicas de predicción son el *core* del modelo predictivo, estas se encargarán de procesar los factores del conjunto de datos ingresados. Para el modelo predictivo propuesto se consideran tres técnicas, las cuales fueron elegidas por la respuesta a la pregunta Q4 planteada en la sección 2.3 y detalladas en el marco teórico (sección 3.2).

Se plantea un modelo híbrido, como se puede observar en la Figura 4-4, el cual consta de tres técnicas de predicción: Máquina de soporte vectorial, Perceptrón multicapa y Random Forest. Estas técnicas realizarán una predicción de acuerdo a los atributos que presenten los clientes, se debe recordar que un cliente puede ser desertor o no desertor en los próximos tres meses siendo los valores asociados “Sí” y “No”, respectivamente.

Cada una de estas técnicas presenta parámetros requeridos para la tarea de predicción, estos parámetros iniciales han sido tomados de la literatura, son los parámetros que mostraron mejores resultados. A continuación se muestran los parámetros para cada técnica.

- Máquina de Soporte Vectorial

Los parámetros requeridos que se proponen usar para esta técnica se muestran a continuación, estos valores han sido tomados de Wang y Niu (2009):

Tabla 4-4. *Configuración propuesta para la técnica SVM*

CONFIGURACIÓN PROPUESTA		
Función Kernel	Parámetro de Kernel	Parámetro C
RBF	0.28	10

Fuente: elaboración propia con base en Wang y Niu (2009)

- Perceptrón Multicapa con back propagation

Los parámetros requeridos que se proponen usar para esta técnica se muestran a continuación, estos valores han sido tomados de Wang y Niu (2009):

Tabla 4-5. *Configuración propuesta para la técnica Perceptrón Multicapa*

CONFIGURACIÓN PROPUESTA					
Número de neuronas en capa oculta	Tasa de aprendizaje	Momento	Número de eras	Número de neuronas en capa de entrada	Número de neuronas en capa de salida
14	0.04	0.5	500	número de factores	2, (desertor y no desertor)

Fuente: elaboración propia con base en Wang y Niu (2009)

- Random Forests

Los parámetros requeridos que se proponen usar para esta técnica se muestran a continuación, estos valores han sido tomados de Keramati *et al*. (2016):

Tabla 4-6.*Configuración propuesta para la técnica Random Forest*

CONFIGURACIÓN PROPUESTA	
Número de árboles	Número de características
20	5

Fuente: elaboración propia con base en Keramati *et al*. (2016)

Estos parámetros son tomados de la revisión de literatura y serán modificados por prueba y error en la elaboración del modelo.

- **Proceso de entrenamiento y validación**

Para obtener un modelo predictivo con las tres técnicas anteriormente mencionadas es necesario realizar dos fases:

- Entrenamiento; se realiza un entrenamiento supervisado mediante el procesamiento de datos para ajustar los parámetros inicialmente ingresados para cada una de las tres técnicas. Al finalizar esta fase obtendremos un modelo entrenado con los parámetros ajustados.
- Validación; se realiza una verificación con los parámetros finales de la fase de entrenamiento. Al finalizar esta fase se obtiene el porcentaje de acierto del modelo predictivo.

La data será dividida en tres partes iguales, de las cuales se utilizarán dos para el entrenamiento y una para la validación.

- **Esquema general del modelo predictivo**

En la Figura 4-4 se presenta una esquematización del modelo de minería de datos planteado, en este se explican las entradas, la utilización de las técnicas y la salida del modelo de minería de datos para la predicción de deserción de clientes en la banca.

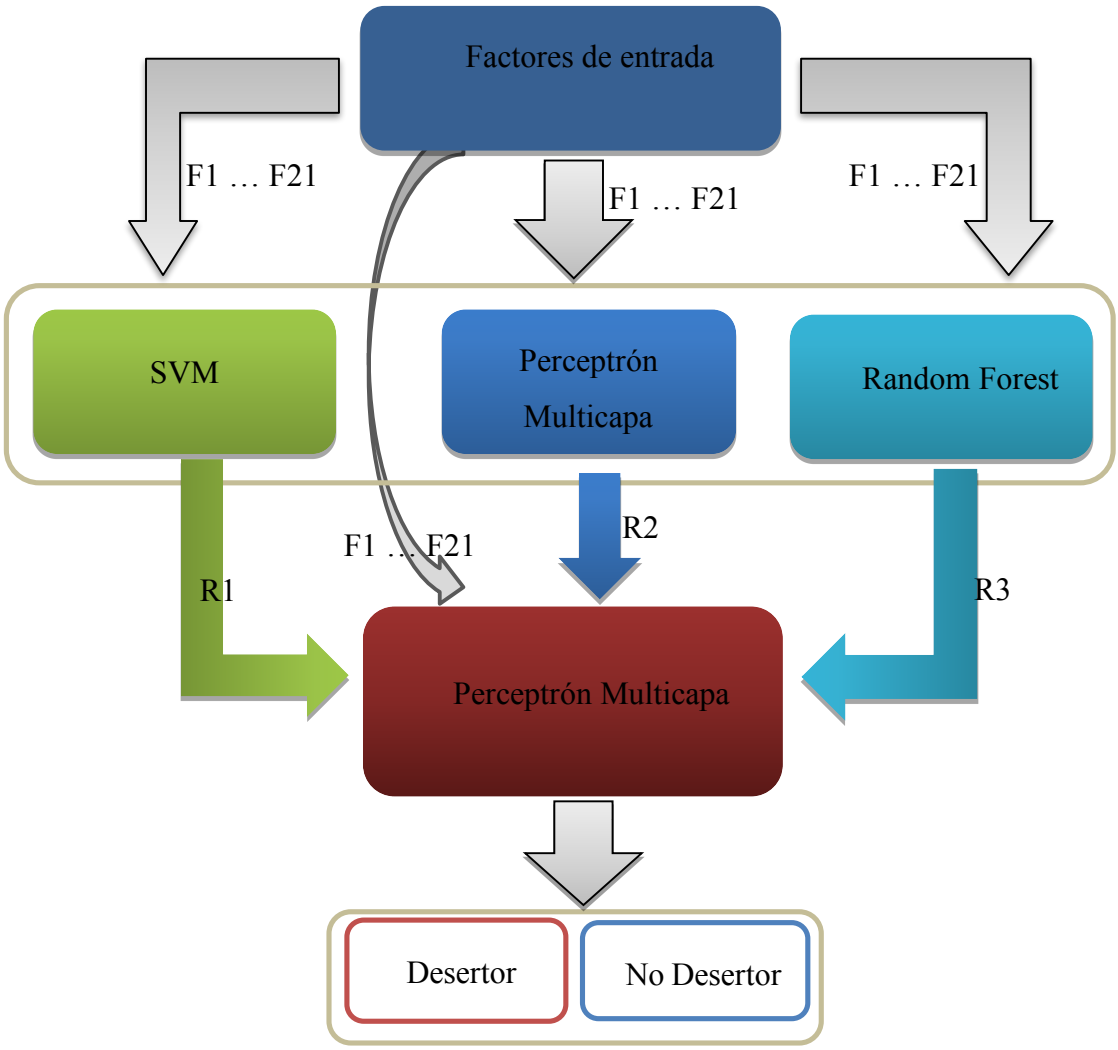


Figura 4-5. Esquema del modelo de minería de datos

Fuente: elaboración propia

Variables de entrada: Los 21 factores definidos en la Tabla 4-2.

Tecnicas seleccionadas (**SVM, Perceptron Multicapa, Random Forest**): Las tecnicas que realizan las primera fase de las prediccion (cada tecnica dara una prediccion por separado), teniendo como entrada los factores definidos anteriormente.

R_1, R_2 y R_3 : salidas obtenidas de cada una de las técnicas de predicción.

La integración de las tres técnicas de predicción será realizada mediante una red neuronal (perceptrón multicapa), la configuración se muestra en la Fuente: elaboración propia.

Tabla 4-7. *Configuración propuesta para el meta-clasificador*

CONFIGURACIÓN PROPUESTA				
Tasa de aprendizaje	Momento	Número de eras	Número de neuronas en capa de entrada	Número de neuronas en capa de salida
0.04	0.5	500	24	2

Fuente: elaboración propia

Esta red neuronal que se usará como meta-clasificador, tendrá por parámetros de entrada los 21 factores definidos en la Tabla 4-2, más las salidas de las 3 técnicas como se muestra en la Tabla 4-8. Este modelo se desarrollara en 2 procesos, el primero sera el entrenamiento de las tecnicas por separadas para obtener la mejor configuración, una vez obtenida se procedera a entrenar al meta-clasificador con las entradas mencionadas anteriormente.

Tabla 4-9. *Factores para el meta-clasificador*

Factores	Descripción	Tipo	Valores Permitidos
	Sociodemográficas		
F1	Edad del cliente	DP	[+18]
F2	Género del cliente	C	[M,F]
F3	Grado de educación del cliente	C	[Primaria, Secundaria, Técnica, Superior]
F4	Ingresos disponibles Mensuales	CP	-
	Percepción de servicio		
F5	Número de reclamos	DP	-
	Comportamiento bancario		
F6	Periodos (Meses) de posesión de tarjeta de crédito	DP	-
F7	Monto del crédito	CP	-
F8	Tiempo en meses de no realizar transacciones	DP	-
F9… F14	Monto de las transacciones en el mes i (1 .. 6)	CP	-
F15… F20	Cantidad de transacciones en el mes i (1 .. 6)	DP	-
F21	Cantidad promedio de las transacciones mensuales	DP	-
	Salida de las tecnicas		

R1	Salida de la tecnica SVM	DP	-
R2	Salida de la tecnica Perceptrón Multicapa	DP	-
R3	Salida de la tecnica Randon Forest	DP	-
	Resultado		
F25	Cliente desertor o no en los próximos tres meses	C	[SI, NO]

Fuente: elaboración propia

Se introduce la propuesta del meta-clasificador debido a que no se puede determinar cuál es la naturaleza de la función que describe la salida y también porque es un método adecuado para ensamblar varios clasificadores (Kuncheva, 2004).

La salida será determinada por lo siguiente.

Si: Y = 0 entonces resultado final “NO”

Y = 1 entonces resultado final “SI”

Esta salida representa si el cliente será desertor o no en los próximos tres meses.

4.5.4 Métricas

La métrica es una medida o el conjunto de ellas para evaluar una determinada característica, en este caso se evaluará el desempeño del modelo de predicción de deserción de clientes en la banca, en específico se evaluará su acierto.

En esta sección se describirá la métrica seleccionada para la evaluación del caso de estudio. Como se pudo observar en la sección 2.3 Resultados, existen diversas métricas para la evaluación de los experimentos realizados que fueron estudiados en los papers seleccionados.

La métrica seleccionada para este estudio es el accuracy (precisión), como se puede observar en la Q4 de la sección 2.3 Resultados es la métrica más utilizada, los 20 papers leídos la emplean.

Precisión: la precisión es la medida de la proporción verdadera positivos y verdaderos negativos, que se identifican correctamente. Se puede expresar matemáticamente de la siguiente manera.

- Tasa de precisión: (A+D)/(A+B+C+D)

Tabla 4-9. *Matriz de clasificación*

Estado del cliente	Predicción de deserción	Predicción de no deserción
Deserción real	A	B
No deserción real	C	D

Fuente: elaboración propia

5 Capítulo 5. Desarrollo del Sistema para la predicción de deserción de clientes en la banca

En este capítulo se describen las características del Sistema de predicción de deserción de clientes en la banca, en donde se indica la arquitectura, modelos, diagramas y roles del sistema.

5.1 Descripción general del sistema

Se desarrollará un sistema que se encargará de realizar como punto principal la predicción de deserción de clientes bancarios y como puntos secundarios el entrenamiento del modelo mediante el ingreso de data de los clientes y de parámetros para las diferentes técnicas de predicción que se utilizarán.

El alcance del Sistema abarca lo siguiente:

- Diseñar un control de acceso para validar los usuarios que ingresan al sistema.
- Generación de modelos mediante el entrenamiento de tres técnicas predictivas (SVM, Random Forest y RN), para ello se ingresaran datos de los clientes y los parámetros de cada una de las técnicas.
- Validación de los modelos mediante el ingreso de datos distintos a los del entrenamiento.
- Predicción de clientes desertores utilizando los modelos creados.

5.2 Arquitectura del Sistema

Debido a que dentro de los requerimientos funcionales del sistema, este debe brindar un ambiente de análisis disponible vía web, se definió la arquitectura Cliente-Servidor, en donde se usó la tecnología Java con JSF (Java Server Faces) con la implementación de PrimeFaces para la interfaz del usuario, MySQL como sistema gestor de base de datos, la utilización de Hibernate para la conexión entre la aplicación y la base de datos, el servidor de aplicaciones Glassfish y la librería WEKA que nos permite la creación de nuestros modelos para la predicción de deserción, la arquitectura se muestra en el Figura 5-1.

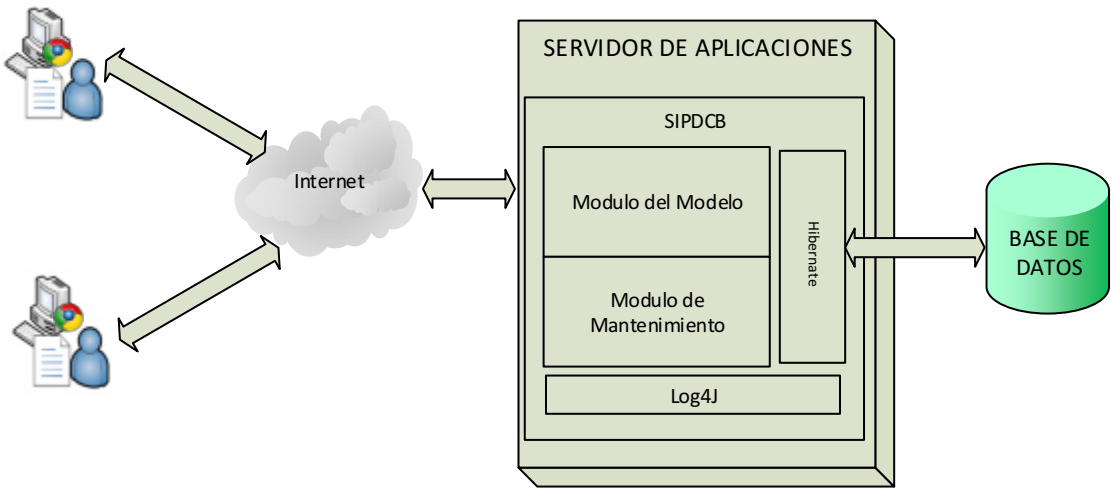


Figura 5-1. Arquitectura del Sistema Web

Fuente: elaboración propia

La principal ventaja de esta arquitectura es que no existen problemas al momento de realizar actualizaciones en la aplicación, dado que es totalmente transparente para los usuarios, y pueden ver de inmediato las modificaciones.

En la Fuente: elaboración propia

se muestra cada uno de los elementos de la arquitectura del sistema.

Tabla 5-1. *Elementos del sistema para la predicción de deserción*

Elemento	Nombre	Especificaciones
JEE Web	Java Enterprise Edition	Versión: 7
Servidor web	GlashFish	Versión: 4.1 Puerto: 8080
Framework de mapeo objeto relacional	Hibernate	Versión: 4.3.1
Gestor de Base de Datos	Mysql	Versión: 5.6 Puerto: 3306
Navegador	Google Chrome	Versión 35.0
Sistema Operativo	Windows 7	64 bits
Framework para desarrollo Web	JSF	JSF 2.2
Framework para minería de datos	Weka	Versión 3.6.10

Fuente: elaboración propia

5.3 Diagrama de componentes

Mediante la Figura 5-2 muestran los componentes que se utilizarán para el desarrollo de cada uno de los módulos que se implementaran para el sistema de predicción de deserción de clientes bancarios.

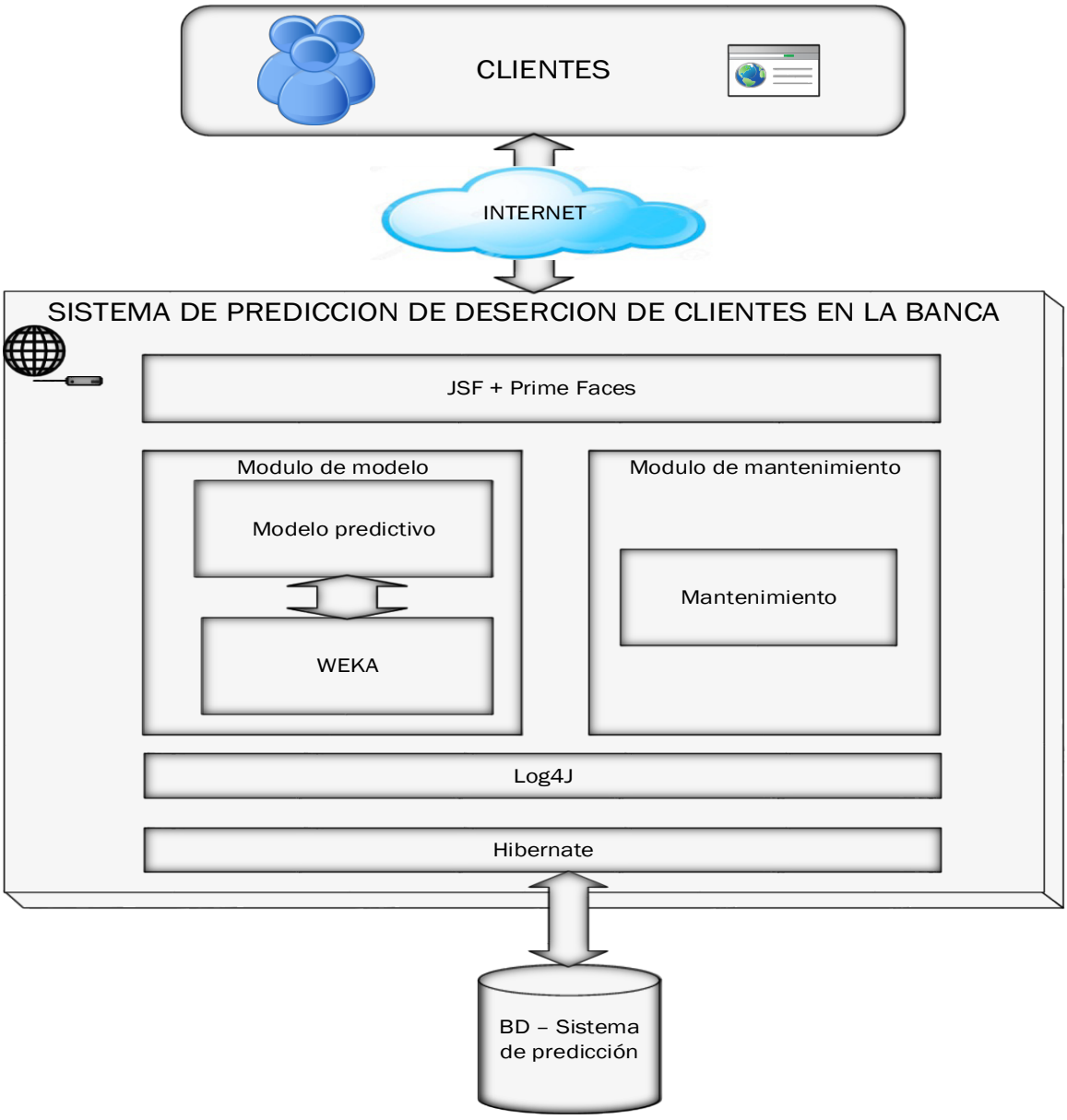


Figura 5-2. Diagrama de componentes del sistema

Fuente: elaboración propia

5.4 Modelado de Casos de Uso del Sistema

Para el modelado del sistema se describe a los usuarios del sistema, los Casos de Uso del Sistema y la representación de las funcionalidades a través de las interfaces.

5.4.1 Usuarios del Sistema

El Sistema admite 2 tipos de usuarios que se describen a continuación:

1. Especialista
    - Descripción: persona encargada de la configuración de parámetros, entrenamiento, validación del modelo y mantenimiento de modelos y técnicas además del mantenimiento de los usuarios.
    - Funciones:
        - Configurar parámetros, previo al entrenamiento de un modelo predictivo, el especialista realiza la configuración de parámetros.
        - Entrenar los modelos para una correcta detección de clientes desertores, es importante pues cada cierto tiempo el comportamiento de los clientes podría variar. También observar la tasa de precisión en la deserción de clientes.
        - Validar el modelo SVM en donde el sistema cargara el modelo previamente entrenado y el usuario ingresa los datos de clientes diferentes al usado para el entrenamiento y en donde se puede observar la tasa de precisión en la predicción de deserción.
        - Creación de nuevos usuarios para el sistema.
2. Usuario
    - Descripción: persona encargada de predecir la deserción de clientes.
    - Funciones:
        - Detectar a los clientes desertores, mediante la carga de clientes a los cuales se desconoce presentan tendencias a ser desertores o no, el sistema usara el modelo entrenado y validado, y mostrará a los clientes que en el futuro serán desertores.

1.1.1 Diagrama de Casos de Uso

Una vez conocido los usuarios del sistema se definen los requerimientos utilizando la técnica de casos de uso. A continuación se muestra el diagrama de casos de uso agrupado en paquetes como se puede ver en la Figura 5-3.

Los paquetes de casos de uso que contiene el sistema son:

- Generación de Modelo
- Mantenimiento
- Proceso de Predicción

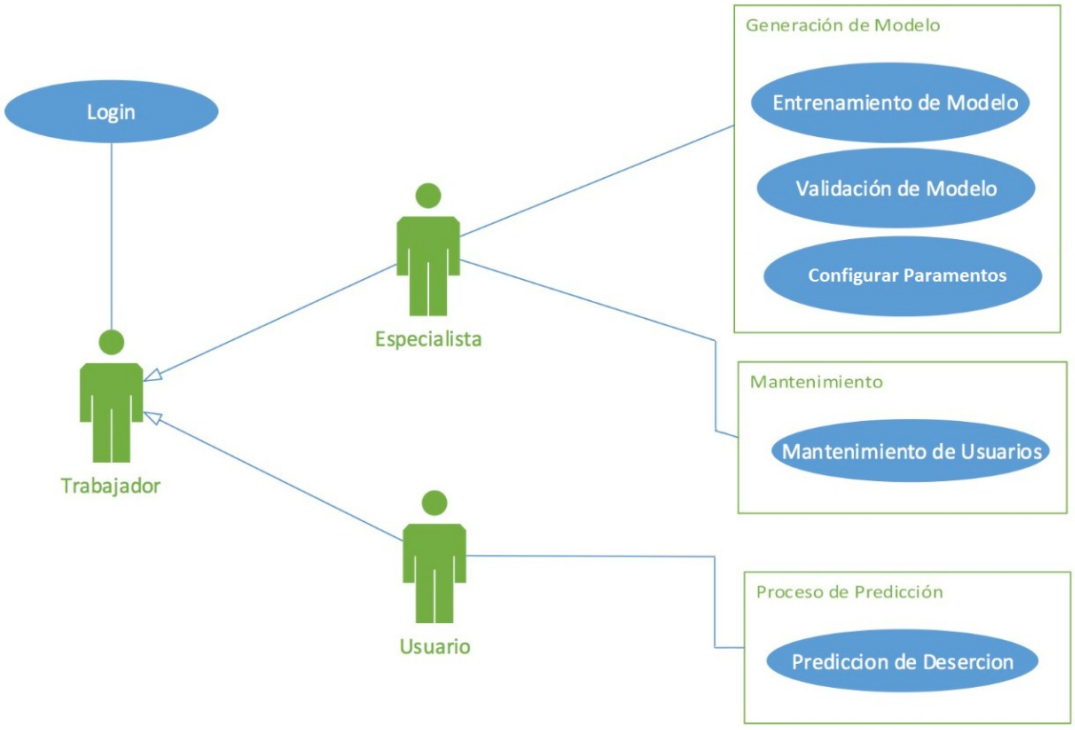


Figura 5-3. Diagrama de casos de uso del sistema

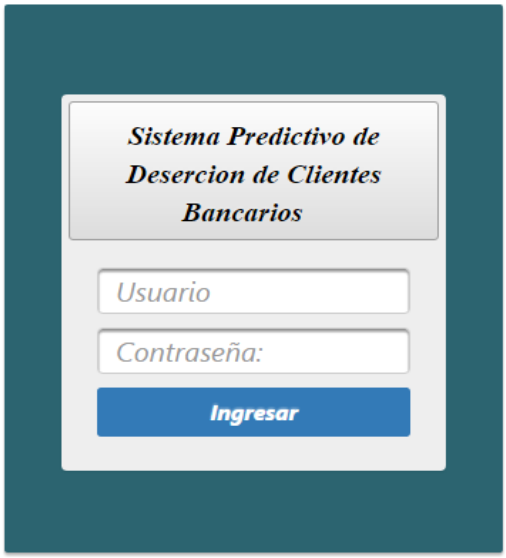
Fuente: elaboración propia

5.4.2 Casos de Uso

A continuación se muestra los casos de uso del sistema:

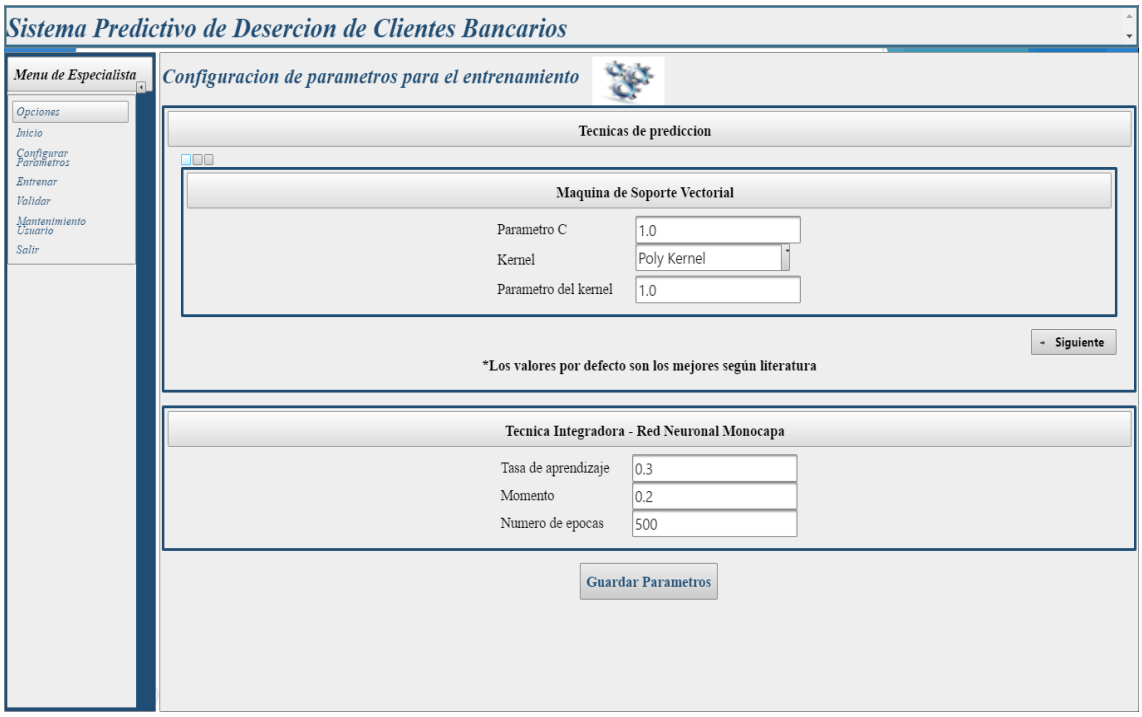
Tabla 5-2. *Casos de usos*

ID:	CUS01
Caso de Uso:	Login
Actor:	Trabajador
Descripción:	Este caso de uso permite validar al usuario del sistema con el identificador y contraseña previamente asignados.
Precondición:	Ninguno.
Flujo Principal:	
1. El caso uso inicia cuando el usuario ingresa al sistema. 2. El sistema muestra un formulario con los siguientes elementos: • Un Input Text para el ingreso del identificador de usuario. • Un Input Text para el ingreso de la contraseña. • Un Botón para realizar la acción del Logeo con el texto "Ingresar". 3. El usuario ingresa el identificador, la contraseña y presiona el botón "Ingresar". 4. El sistema valida la existencia de datos ingresados en los campos del identificador y la contraseña. 5. El sistema: 5.1. Busca el identificador de usuario. 5.2. Compara la contraseña ingresada con la del identificador encontrado. 5.3. Lee el perfil asociado al identificador del usuario. 5.4. Asigna los permisos para ingresar al sistema. 5.5. Muestra la pantalla de menú inicial según el perfil del usuario. 6. El usuario ingresa al sistema con el perfil asignado y el caso de uso finaliza	
Post-condición :	El usuario ha sido validado por el sistema y puede acceder al sistema con el perfil asociado.
Flujo Alterno:	"Identificador o Contraseña incorrecta"
1. En el paso 5 el sistema encuentra que los datos del login son incorrectos, el sistema muestra el siguiente mensaje: "Identificador y/o Contraseña incorrecto" y el caso de uso continua el paso 2.	
Prototipo:	



ID:	CUS02
Caso de Uso:	Configurar Parámetros
Actor:	Especialista
Descripción:	Este caso de uso permite configurar los parámetros de las técnicas usadas para el modelo de predicción
Precondición:	El usuario debe ser admitido por el sistema y contar con el perfil de Especialista.
Flujo principal:	
1. El caso de uso comienza cuando el especialista selecciona la opción “Configurar Parámetros”. 2. El sistema muestra un formulario con los siguientes elementos: • Una sección con campos correspondientes a los parámetros de cada técnica de predicción • Un Botón para elegir la siguiente Técnica. • Una sección con campos correspondientes a los parámetros de la técnica integradora. • Un Botón para “Guardar Parámetros”. 3. El especialista ingresa los valores correspondientes a la primera técnica y presiona siguiente. 4. El sistema muestra la sección correspondiente de la segunda técnica. 5. El especialista ingresa los valores correspondientes a la segunda técnica y presiona siguiente. 6. El sistema muestra la sección correspondiente de la tercera técnica. 7. El especialista ingresa los valores correspondientes a la tercera técnica. 8. El especialista ingresa los valores correspondientes a la técnica integradora y da clic en el “Guardar Parámetros”. 9. El sistema guarda la configuración y muestra el siguiente mensaje “Se guardaron los parámetros de configuración”	
Post-condición :	El especialista ha realizado la configuración de los parámetros correspondiente a las técnicas satisfactoriamente.

Flujo Alterno:	“Parámetros por defecto”
1. En el paso 3, 5 o 7 el usuario presiona el botón “Guardar Paramentos”, el cus continua en el paso 9.	
Prototipo:	



ID:	CUS03
Caso de Uso:	Entrenamiento de Modelo
Actor:	Especialista
Descripción:	Este caso de uso permite realizar el entrenamiento de un nuevo modelo para la predicción de deserción.
Precondición:	El usuario debe ser admitido por el sistema y contar con el perfil de Especialista.
Flujo principal:	

1. El caso de uso comienza cuando el especialista selecciona la opción “Entrenar”.
2. El sistema muestra un formulario con los siguientes elementos:

- Un botón para realizar la búsqueda del archivo CSV que contiene la información de los clientes, con los que se realizara el entrenamiento.
- Un botón para eliminar del sistema el archivo seleccionado.
- Una sección donde se muestran los parámetros por defecto de cada técnica de predicción (En cuadros de texto editables).
- Un Input Text para ingresar el nombre del modelo que se entrenará (Por defecto muestra un nombre compuesto por “modelo_%fecha%-%hora%”).
- Una sección que muestra el porcentaje de error del entrenamiento de cada técnica.
- Un botón para realizar el entrenamiento con el texto “Entrenar Modelo”.

3. El usuario presiona el botón para realizar la búsqueda del archivo y lo selecciona
4. El sistema verifica que el archivo seleccionado sea CSV, posteriormente carga el archivo.
5. El usuario presiona el botón “Entrenar modelo”.
6. El sistema:

6.1. Realiza el entrenamiento con el archivo CSV, los parámetros de cada técnica y la función para el resultado final.

6.2. Guarda el modelo en base de datos

6.3. Muestra un popup con el mensaje “Se entrenaron y guardaron los modelos.” y completa la sección con los porcentajes de error del entrenamiento.

7. El caso de uso finaliza.

Post-condición :	El especialista ha realizado el entrenamiento del modelo para la predicción de deserción.
Flujo alterno:	“Tipo de archivo incompatible”

2. En el paso 4, el sistema verifica que el archivo no es de tipo CSV por lo que muestra el mensaje: “El archivo no es de tipo CSV, seleccione un archivo de este tipo.” Y el caso de uso continúa en el paso 3.

Flujo Alterno:	“No se cargó archivo”

En el paso 3 el usuario presiona el botón “Entrenar Modelo” y el sistema identifica que no se ha seleccionado ningún archivo para el entrenamiento, por lo cual muestra el mensaje “Necesitas cargar un archivo de entrenamiento” y el caso de uso continúa en el paso 3.

Prototipo:	

ID:	CUS04
Caso de Uso:	Validación de Modelo
Actor:	Especialista
Descripción:	Este caso de uso permite realizar la validación de un modelo para la predicción de deserción.
Precondición:	El usuario debe ser admitido por el sistema y contar con el perfil de Especialista.
Flujo Principal:	

1. El caso de uso comienza cuando el usuario selecciona la opción “Validar”.
2. El sistema muestra una pantalla para la selección del modelo entrenado con los siguientes elementos:
   - Un botón para realizar la búsqueda del archivo CSV que contiene la información de los clientes, con los que se realizara la validación.
   - Un botón para eliminar del sistema el archivo seleccionado.
   - Un campo auto-completable para buscar el modelo por el nombre.
   - Una sección que muestra el porcentaje de acierto de la validación de cada técnica.
   - Un botón para realizar la validación, con el texto “Validar”.
3. El usuario presiona el botón para realizar la búsqueda del archivo y lo selecciona.
4. El sistema verifica que el archivo seleccionado sea CSV, posteriormente carga el archivo.
5. El sistema verifica que el archivo seleccionado sea CSV, posteriormente carga el archivo.
6. El usuario ingresa el nombre del modelo.
7. El sistema realiza el filtrado por el nombre ingresado.
8. El usuario selecciona el modelo de la lista de modelos filtrados y presiona el botón “Validar”.
9. El sistema:
   9.1. Realiza la validación con el archivo CSV.
   9.2. Actualiza el modelo en base de datos.
   9.3. Completa la sección de los porcentajes de la validación.
   9.4. Muestra un mensaje “Modelo Validado”
10. El caso de uso finaliza.

Post-condición:	El especialista ha realizado la validación del modelo para la predicción de deserción.

Flujo Alterno:	“No se cargó archivo”

1. En el paso 3 el usuario presiona el botón “Validar” y el sistema identifica que no se ha seleccionado ningún archivo para la validación, por lo cual muestra el mensaje “Necesitas cargar un archivo de validación” y el caso de uso continúa en el paso 3.

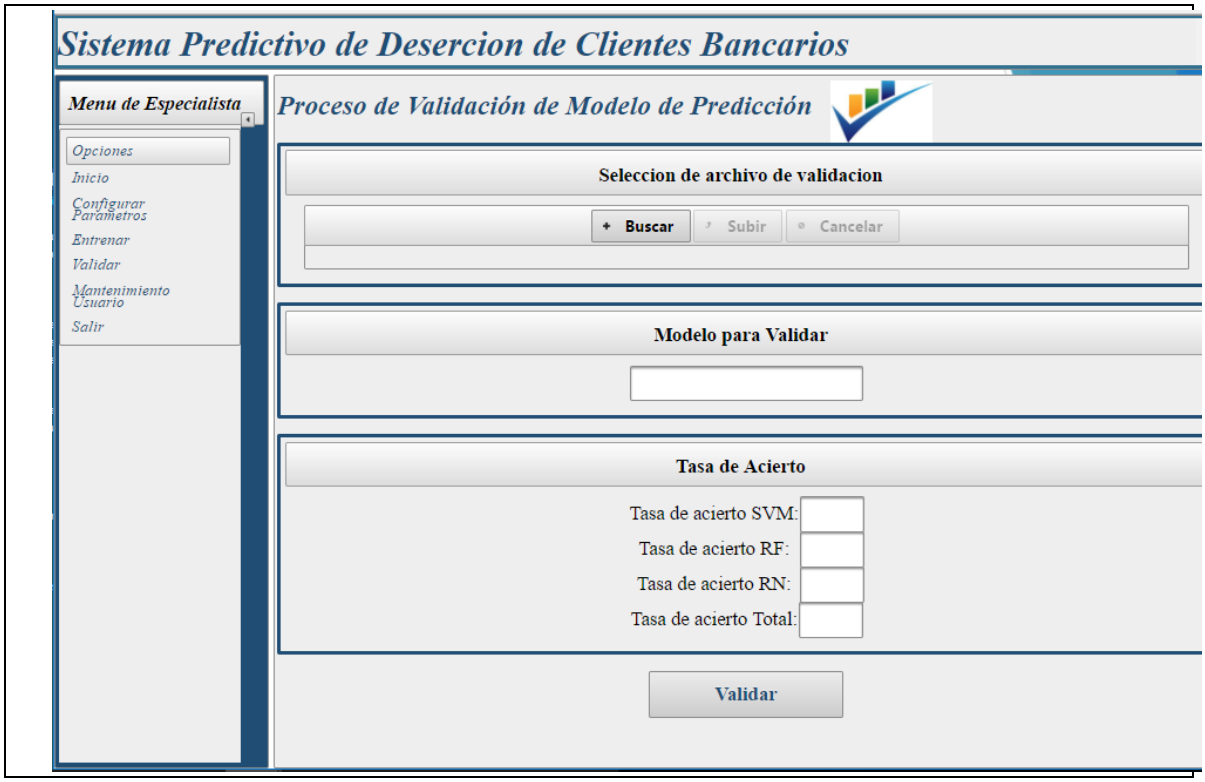
Flujo Alterno:	“Modelo no seleccionado”

1. En el paso 5 el usuario presiona el botón “Validar” y el sistema identifica que no se ha seleccionado ningún modelo para la validación, por lo cual muestra el mensaje “Necesitas cargar un modelo para la validación” y el caso de uso continúa en el paso 3.

Flujo Alterno:	“Archivo incompatible”

1. En el paso 5, el sistema verifica que el archivo no es de tipo CSV por lo que muestra el mensaje: “El archivo no es de tipo CSV, seleccione un archivo de este tipo.” Y el caso de uso continúa en el paso 7.

Prototipo:	



ID:	CUS05
Caso de Uso:	Predicción de deserción
Actor:	Usuario
Descripción:	Este caso de uso permite realizar la predicción de deserción de clientes de la banca.
Precondición:	El usuario debe ser admitido por el sistema y contar con el perfil de Usuario.
Flujo principal:	

1. El caso de uso comienza cuando el usuario selecciona la opción “Realizar Predicción”.
2. El sistema muestra un formulario con los siguientes elementos:
   - Un botón para realizar la búsqueda del archivo CSV que contiene la información de los clientes, con los que se realizará la validación.
   - Un botón para eliminar del sistema el archivo seleccionado.
   - Un campo de texto para buscar el modelo por el nombre.
   - Un Input Text para ingresar el nombre del proceso de predicción (Por defecto muestra un nombre compuesto por “ProcesoPredictivo_%fecha%-%hora%”).
   - Un botón para realizar la predicción con el texto “Realizar Predicción”.
   - Un cuadro de texto para mostrar la cantidad de clientes desertores (inicialmente oculto).
   - Un cuadro de texto para mostrar la cantidad de clientes no desertores (inicialmente oculto).
   - Una tabla con las características más significativas de todos los clientes desertores (inicialmente oculto).
3. El usuario presiona el botón para realizar la búsqueda del archivo, lo selecciona.
4. El sistema verifica que el archivo seleccionado sea CSV, posteriormente carga el archivo.
5. El usuario ingresa el nombre del modelo.
6. El sistema realiza el filtrado por el nombre ingresado.
7. El usuario selecciona el modelo de la lista de modelos filtrados, ingresa el nombre del proceso de predicción y presiona el botón “Realizar predicción”.
8. El sistema:
   8.1. Realiza el proceso predictivo con el archivo CSV y el modelo seleccionado.
   8.2. Guarda los datos del proceso predictivo en base de datos.
   8.3. Guarda los clientes desertores en base de datos.
   8.4. Muestra la cantidad de clientes desertores y no desertores.
   8.5. Muestra en la tabla cada uno de estos clientes con sus características más relevantes.
   8.6. Muestra un popup con el mensaje “Proceso predictivo realizado”.
9. El caso de uso finaliza.

Post-condición:	El trabajador ha realizado el proceso de predicción de clientes de la banca.
Flujo Alterno:	“Archivo incompatible”

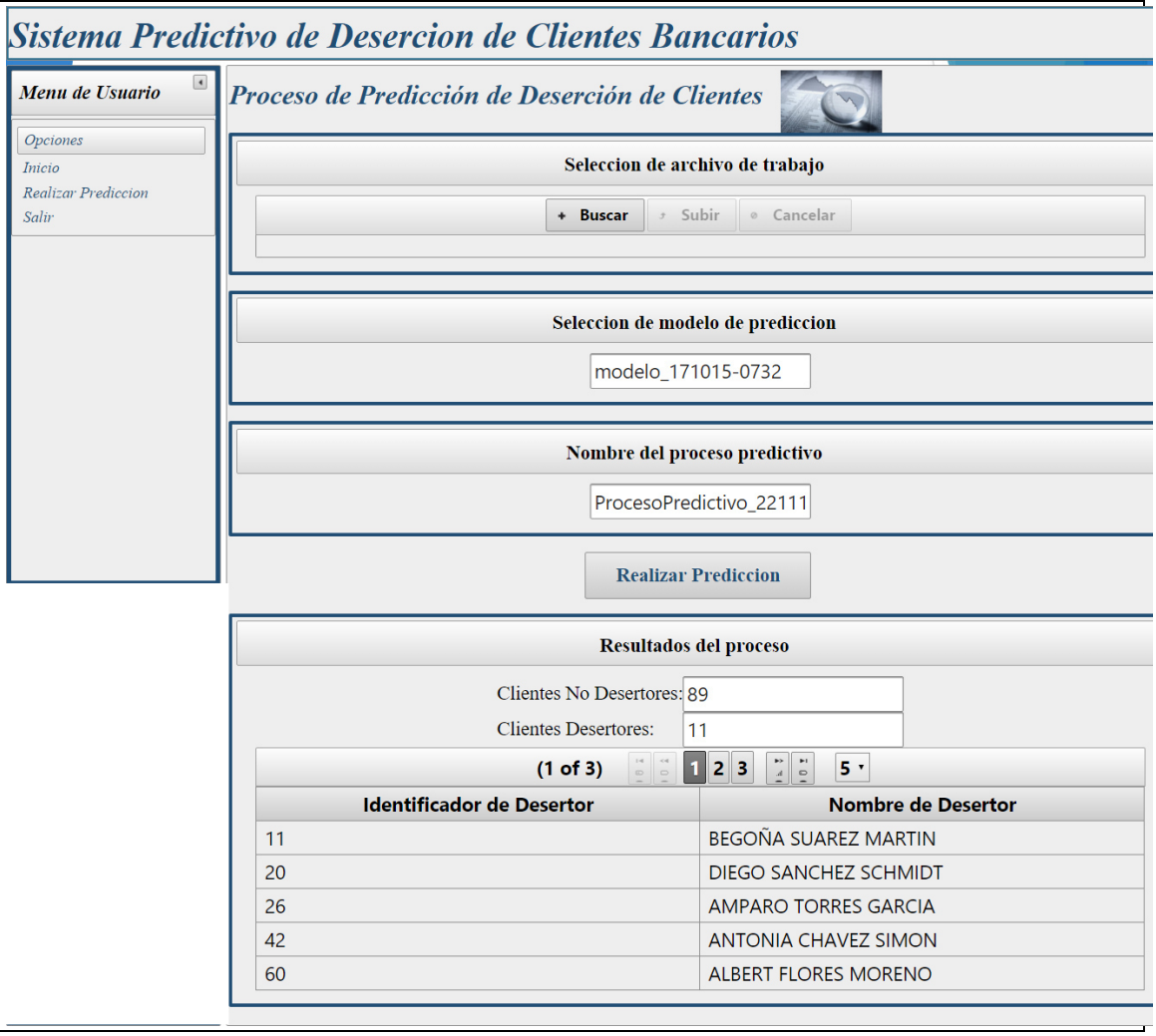
1. En el paso 4, el sistema verifica que el archivo no es de tipo CSV por lo que muestra el mensaje: “El archivo no es de tipo CSV, seleccione un archivo de este tipo.” Y el caso de uso continúa en el paso 3.

Flujo Alterno:	“No se cargó archivo”

1. En el paso 3 el usuario presiona el botón “Realizar Predicción” y el sistema identifica que no se ha seleccionado ningún archivo para la validación, por lo cual muestra el mensaje “Necesitas cargar un archivo de predicción” y el caso de uso continúa en el paso 3.

Flujo Alterno:	“Modelo no seleccionado”

1. En el paso 5 el usuario presiona el botón “Realizar Predicción” y el sistema identifica que no se ha seleccionado ningún modelo para la validación, por lo cual muestra el mensaje “Necesitas cargar un modelo para la predicción” y el caso de uso continúa en el paso 3.
Prototipo:



ID:	CUS05
Caso de Uso:	Mantenimiento de usuarios
Actor:	Especialista
Descripción:	Este caso de uso permite realizar el mantenimiento de los modelos.
Precondición:	El usuario debe ser admitido por el sistema y contar con el perfil de Especialista.
Flujo principal:	

1. El caso de uso comienza cuando el usuario selecciona la opción “Mantenimiento Usuario”.
2. El sistema muestra un formulario con los siguientes elementos:
   - Un botón para agregar un nuevo usuario, con el texto “Agregar”.
   - Una tabla de los usuarios existentes con las siguientes columnas para cada uno: “Usuario” (además de una caja de texto para el filtrado) y “Perfil” (por defecto mostrará todos los usuarios hasta el momento).
   - Botones para la paginación de la tabla de modelos (Primero, Anterior, Siguiente, Ultimo).
3. El usuario presiona el botón agregar.
4. El sistema muestra el siguiente formulario:
   - Un campo de texto para ingresar el nombre del usuario.
   - Un campo de texto para ingresar la contraseña.
   - Una lista desplegable para seleccionar el perfil del usuario.
   - Un botón para registrar el usuario, con el texto “Guardar”.
5. El especialista ingresa el nombre del usuario, la contraseña y el perfil del mismo, posteriormente presiona el botón “Guardar”.
6. El sistema:
   6.1. Realiza la verificación de que el usuario a registrar no exista.
   6.2. Guarda en base de datos los datos del usuario.
   6.3. Actualiza la tabla de los usuarios con la información del nuevo registro.
7. El caso de uso finaliza.

Post-condición:	El Especialista ha realizado mantenimiento de un modelo.
Flujo Alterno:	“Usuario ya existente”

1. En el paso 6.1 el sistema detecta que ya existe un usuario con ese nombre entonces muestra el mensaje “Usuario ya registrado” y el caso de uso continúa en el paso 5.

Prototipo:

5.5 Modelo de datos

El modelo de base de datos será utilizado para la permanencia de datos importantes en el sistema de predicción de deserción de clientes, este modelo se muestra en la Figura 5-4.

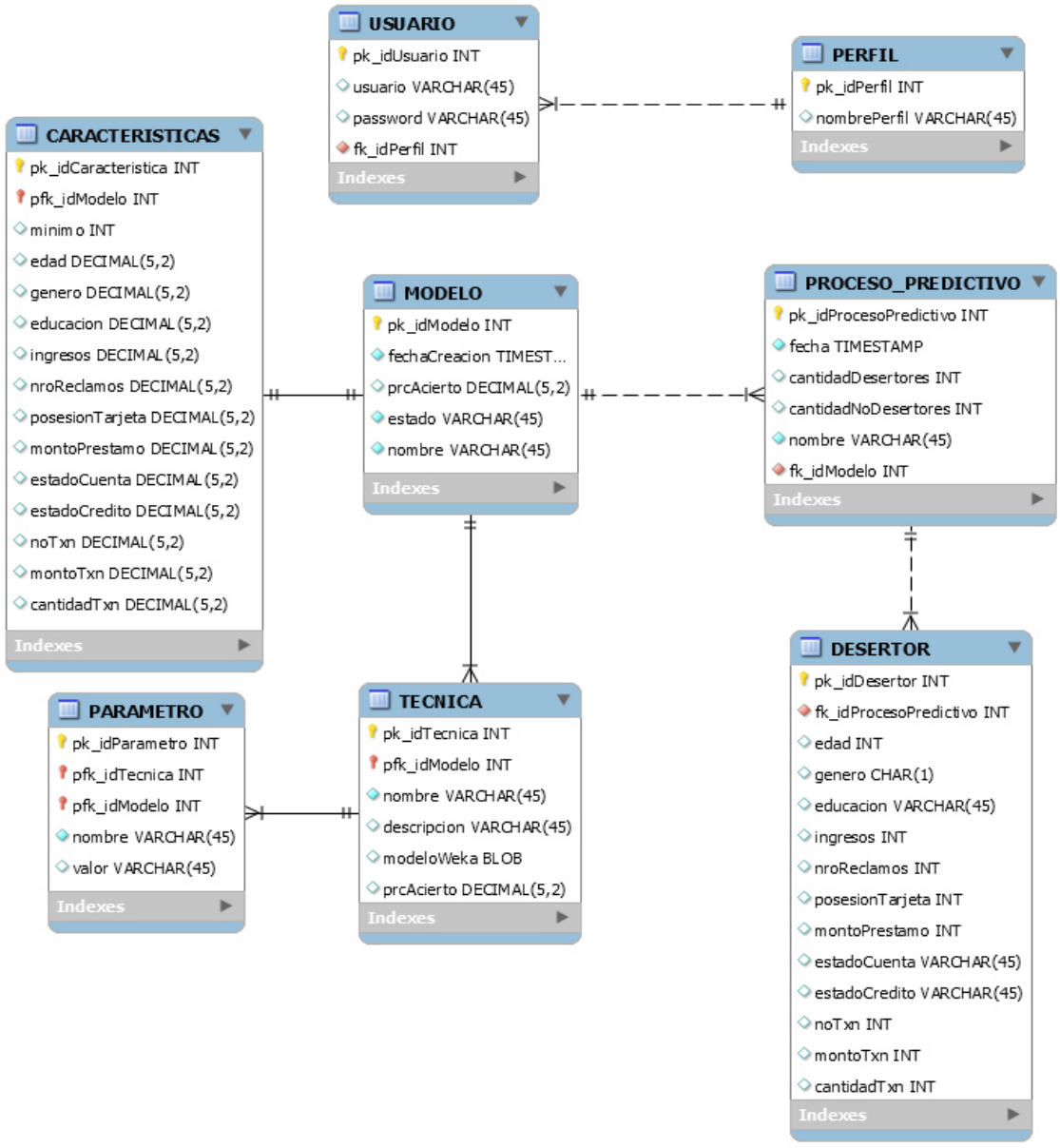


Figura 5-4. Modelo de Datos del sistema para la predicción de la deserción de clientes en la banca

Fuente: elaboración propia

6 Capítulo 6. Pruebas y validación

En este capítulo se evalúa si el modelo construido en las fases anteriores de este estudio cumple con lo esperado, es decir, el sistema realiza una predicción para la deserción de clientes bancarios con un margen de error aceptable.

6.1 Planificación de la validación

Se realizó un plan de validación, como se puede observar en Figura 6-1, el cual se describe a continuación:

Como primera instancia se entrena el modelo de predicción de deserción de clientes bancarios (con los parámetros encontrados en la literatura y la data de entrenamiento), posteriormente se realiza la validación del modelo predictivo (con la data de validación), en caso de que la validación no muestre los resultados esperados se procede a ajustar los parámetros de las técnicas, para volver al punto de entrenamiento del modelo de predicción y repetir el ciclo hasta que el criterio de aceptación (precisión) sea cubierto.

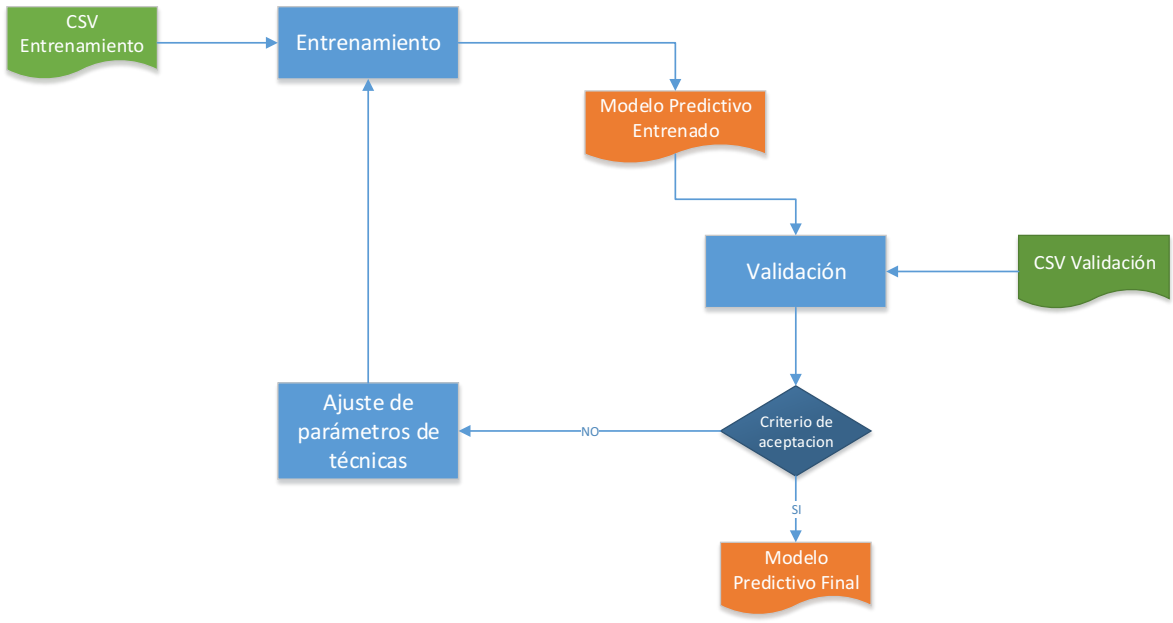


Figura 6-1. Plan de validación del modelo

Fuente: elaboración propia

6.2 Conjunto de datos

Los conjuntos de datos usados por los autores no están disponibles debido a que la data utilizada contiene información sensible de los clientes bancarios, por lo que para realizar las pruebas se utilizó un conjunto de datos que está conformado por 24 420 registros de un banco peruano, el dataset está enfocado en clientes de la ciudad de Lima, de los cuales el 44.76% (10931) son mujeres y el 55.24% (13489) son hombres, además se evidencia que el 15.80% (3859) se encuentra entre las edades de 23 a 35 años, el 54.78% (13378) entre las edades de 35 a 50 años y el 29.42% (7183) entre las edades de 50 a 65 años.

El dataset utilizado es el segundo más grande comparado con los mostrados en la literatura (Fuente: elaboración propia

), siendo el primero el de López et al. (2016) con 52398 registros.

Del conjunto de datos se tomó el 95% para el proceso de entrenamiento, mientras que el 5% restante fue destinada a la validación del modelo predictivo. Por otro lado, se conoce que la cantidad de clientes desertores y los clientes no desertores son 9.71% versus un 90.28% respectivamente. En la Fuente: elaboración propia

-1 se muestra un resumen del conjunto de datos.

Tabla 6-1. *Resumen de conjunto de datos*

	Registros	Desertores (9.71%)	No Desertores (90.28%)
Entrenamiento (95%)	23 199	2 253	20 946
Validación (5%)	1 221	119	1 102
Total	24 420	2 372	22 048

Fuente: elaboración propia

6.3 Ejecución de la validación y ajuste de parámetros

La ejecución de las pruebas se realizó de la siguiente manera:

Como resultado de la implementación del modelo propuesto planteado en el 4y 5se obtiene el modelo entrenado según los parámetros encontrados en la literatura, siguiendo con el proceso de KDD se procedió a la validación del modelo, utilizando el 5% del conjunto de datos.

6.3.1 Validación de factores

Para la validación de los dos factores propuestos en la sección 4.4 se utilizó el coeficiente de correlación de Pearson, para lo cual se usaron los 23 199 registros del dataset de entrenamiento y esta correlación se evaluó entre el factor propuesto y la variable final (desertor o no desertor).

Del experimento se obtuvieron los siguientes resultados:

Tabla 6-2. *Resultados de coeficiente de correlación de Pearson*

	Monto de consumo promedio	**Cantidad de transacciones promedio**
Coeficiente de correlación	-0.124030	-0.557030
Interpretación	Relación inversa y débil	Relación inversa y fuerte

Fuente: elaboración propia

De los resultados mostrados en la Fuente: elaboración propia

se identificó que solo la inclusión del factor "Cantidad de transacciones promedio" es favorable para el experimento, mientras que el factor "Monto de consumo promedio" queda descartado.

6.3.2 Validación y prueba de parámetros

En la Tabla 6-3, Tabla 6-4 y Tabla 6-5 se muestra la mejor configuración y los resultados para cada una de las tres técnicas utilizadas en el modelo propuesto, estas configuraciones se obtuvieron de la literatura.

Tabla 6-6. *Configuración y resultado de SVM con parámetros de la literatura*

	SVM			Entrenamiento	Validación
	Función Kernel	Parámetro Kernel	Parámetro C		
Literatura	RBF	0.28	10	98.3%	92.27%

Fuente: elaboración propia

Tabla 6-7. *Configuración y resultado de RN con parámetros de la literatura*

	Red Neuronal Perceptrón Multicapa						Entrenamiento	Validación
	Neuronas en capa oculta	Tasa de aprendizaje	Momento	Número de eras	Neuronas en la capa de entrada	Neuronas en la capa de salida		
Literatura	14	0.04	0.5	500	21	2	96.8%	82.67%

Fuente: elaboración propia

Tabla 6-8. *Configuración y resultado de RF con parámetros de la literatura*

	Random Forest		Entrenamiento	Validación
	Número de Árboles	Número de características		
Literatura	50	5	100%	92.42 %

Fuente: elaboración propia

En la Fuente: elaboración propia

se indican los parámetros inicialmente propuestos para la técnica integradora.

Tabla 6-9. *Configuración y resultado de Percerptrón Multicapa*

	Red Neuronal Perceptrón Multicapa					Entrenamiento	Validación
	Neuronas de entrada	Tasa de aprendizaje	Momento	Número de eras	Neuronas en la capa de salida		
Parámetros iniciales	25	0.04	0.5	500	2	98.9%	93.72%

Fuente: elaboración propia

Debido a que luego de la validación no se obtuvieron resultados mejores a los de la literatura, se realizó un ajuste de los parámetros para obtener una mejor tasa de precisión, y los resultados no fueron del todo satisfactorios. Posteriormente se aplicaron varios ajustes a los parámetros hasta obtener resultados favorables. El resumen de estos ajustes y las tasas de validación de los principales resultados se muestran en la Tabla 6.7, Tabla 6.8, Tabla 6.9 y Tabla 6.10

.

Tabla 6-10. *Comparación de resultados técnica SVM con ajuste de parámetros*

	SVM			Entrenamiento	Validación
	Función Kernel	Parámetro Kernel	Parámetro C		
Literatura	RBF	0.28	10	98.36%	92.27%

Prueba 1	RBF	0.25	10.5	80.7%	94.45%
Prueba 2	RBF	0.22	11	83.94%	81.67%
Prueba 3	RBF	0.34	11.5	95.32%	74.60%
Prueba 4	RBF	0.44	11	100%	70.67%
Prueba 5	RBF	0.26	10.5	99.65%	97.21%
Prueba 6	Polinomial	0.43	9.5	95.7%	82.31%
Prueba 7	Polinomial	0.29	10	98.38%	92.01%
Prueba 8	Polinomial	0.35	12	100%	86.42%
Prueba 9	Polinomial	0.31	10.5	100%	93.26%
Prueba 10	Polinomial	0.31	10	87.4%	91.37%

Fuente: elaboración propia

Tabla 6-11. *Comparación de resultados técnica RN con ajuste de parámetros*

	Redes Neuronales						Entrenamiento	Validación
	Neuronas en capa oculta	Tasa de aprendizaje	Momento	Número de eras	Neuronas en la capa de entrada	Neuronas en la capa de salida		
Literatura	14	0.04	0.5	500	21	2	96.81%	82.67%
Prueba 1	16	0.1	0.5	500	21	2	100%	90.67%
Prueba 2	17	0.08	0.5	800	21	2	100%	92.67%
Prueba 3	17,7	0.05	0.5	1000	21	2	99.61%	96.8%
Prueba 4	16,8	0.12	0.5	1000	21	2	100%	75.54%
Prueba 5	16	0.15	0.5	750	21	2	90.83%	90.67%
Prueba 6	15	0.09	0.5	750	21	2	89.7%	64.54%
Prueba 7	14	0.06	0.5	800	21	2	80.3%	70.76%
Prueba 8	17,9	0.07	0.5	500	21	2	79.84%	64.37%
Prueba 9	16,6	0.1	0.5	800	21	2	100%	91.62%
Prueba 10	14,5	0.12	0.5	1500	21	2	100%	54.67%

Fuente: elaboración propia

Tabla 6-12. *Comparación de resultados técnica RF con ajuste de parámetros*

	Random Forest		Entrenamiento	Validación
	Número de Árboles	Número de características		
Literatura	50	5	100%	92.42%
Prueba 1	40	6	99.64%	76.00%
Prueba 2	48	7	100%	93.11%
Prueba 3	63	7	99.8%	91.21%
Prueba 4	55	6	99.23%	88.87%
Prueba 5	47	5	99.8%	86.39%
Prueba 6	65	5	100%	90.53%
Prueba 7	40	6	99.3%	86.89%

Prueba 8	58	7	100%	97.37%
Prueba 9	58	6	98.63%	90.56%
Prueba 10	65	7	99.7%	92.13%

Fuente: elaboración propia

Como se explicó en Capítulo 4. Modelo propuesto para la predicción de deserción de clientes en la banca 3 de las 24 entradas son las salidas de las 3 técnicas seleccionas, para el ajuste de los parámetros del **Modelo Hibrido** se tomaron las mejores configuraciones de cada una de las técnicas predictivas obtenidas en Tabla 6-6, Tabla 6-7 y Tabla 6-8; con ellas se realizaron 5 pruebas para encontrar la mejor configuración del modelo híbrido.

Tabla 6-13. *Comparación de resultado de Modelo Hibrido.*

	Red Neuronal Perceptrón Multicapa					Entrenamiento	Validación
	Neuronas en capa de entrada	Tasa de aprendizaje	Momento	Número de eras	Neuronas en la capa de salida		
Prueba 1	24	0.04	0.5	500	2	98.82%	96.3%
Prueba 2	24	0.06	0.5	1000	2	98.1%	94.5%
Prueba 3	24	0.07	0.4	500	2	99.4%	96.4%
Prueba 4	24	0.05	0.4	750	2	100%	97.38%
Prueba 5	24	0.11	0.5	500	2	98.96%	92.1%

Fuente: elaboración propia

Luego de realizar las pruebas para ajustar los parámetros de las técnicas predictivas y el modelo híbrido se obtuvieron las siguientes tasas de acierto.

Tabla 6-14. *Resultados finales del ajuste y validación del modelo híbrido*

	SVM	RN	RF	Modelo Híbrido
Mejores Resultados	97.21 %	96.8%	97.37%	97.38%

Fuente: elaboración propia

Los resultados de la Tabla 6-14 son los resultados finales de la validación y son los parámetros que proporcionan estos resultados los que conforman la configuración final del modelo de predicción de deserción de clientes, con esta configuración según el modelo propuesto es posible determinar si un cliente va a ser desertor o no en los siguientes tres meses.

7 Capítulo 7. Conclusiones y trabajos futuros

7.1 Discusión

En el estudio realizado se obtuvieron resultados que mejoran a los presentados en la literatura hasta el momento. El autor Farquad et al. (2014) con una tasa de acierto de 97.10%, (la mejor en el estado del arte), utilizo múltiples técnicas de inteligencia artificial en su modelo predictivo, sin embargo las utilizo en diferentes fases, utilizo la técnica SVM-RFE para la reducción de los atributos en el dataset, posteriormente utilizo SVM para realizar la predicción y finalmente utilizo Naive Bayes Tree para obtener reglas que explicaran el comportamiento de los desertores. Así mismo la segunda mejor tasa de acierto, presentada por Keramati *et al*. (2016), con una tasa de acierto de 96.70%, solo utiliza la técnica Arboles de Decisión para realizar la predicción. En comparación con el trabajo presentado en el cual solo teniendo en cuenta la fase de predicción se combinan las tres mejores técnicas de la literatura: SVM, DT y Redes Neuronales, utilizando un Perceptron Multicapa con la finalidad de mejorar la tasa de acierto presentada en la literatura, objetivo que se logró con un 97.38% en la tasa de acierto.

Un componente determinante en el modelo predictivo son los factores utilizados, en la Tabla 2- 3 se puede observar que aquellos autores que introdujeron nuevos atributos relacionados con el comportamiento bancario obtuvieron una mejor tasa de acierto, por ejemplo: Nie et al. 2011, quien introdujo el factor “quejas de cliente” obtuvo un 84%, Zhu et al. 2014, considerando el factor “monto utilizado” llego a obtener un 92%. En base a estos estudios decidimos utilizar el factor “número de quejas de cliente”, “cantidad de transacciones” y “monto de transacciones” en los últimos seis meses.

Tabla 7-1. *Datasets en el estado del arte con más de 20 000 registros*

Dataset	Acierto	Referencia
20000	78.10%	(Xie *et al*., 2009)
20000	62.00%	(Xie *et al*., 2009)
20000	87.20%	(Xie *et al*., 2009)
52398	67,70%	(López et. al,2016)
20000	93.20%	(Xie *et al*., 2009)
21000	91.60%	(Sin *et al*., 2011)

Fuente: elaboración propia

Este resultado pueden ser comparables con los estudios que cuentan con un data set de más de 20 000 registros como se muestra en la Tabla 7-1, debido a que el data set utilizado en este estudio contiene 24 420 registros. En el presente estudio la tasa final de acierto es de 97.38%, la cual supera a los estudios con dataset similar.

7.2 Conclusiones

Se implementó un sistema inteligente que mediante características y comportamiento de un cliente bancario de micro-crédito de la ciudad de Lima logra predecir la deserción de clientes bancarios para los próximos tres meses, con una eficiencia mejor a la de la literatura. Este sistema inteligente se basa en un modelo híbrido el cual fue detallado en el Capitulo **4**.

Se revisó la literatura e identificaron 21 características más resaltantes para la solución de este problema, dichas características han sido mostradas en la sección **4.5.1**, en la cual se indican y describen todas las variables que influyen en el proceso de creación, ajuste y utilización del modelo.

También se encontraron en la literatura las técnicas empleadas en el problema de la deserción de clientes bancarios. Se pudo identificar que existe una gran variedad de técnicas, las cuales son mencionadas en la sección **2.3 Resultados**.

De igual manera se identificó el modelo más adecuado para el tratamiento del problema de deserción de clientes bancarios, este modelo fue resultado de la revisión de las diferentes

maneras de abordar el problema. Con la ayuda la técnica de minería de datos KDD este modelo fue realizado y probado.

Finalmente, una vez desarrollada la validación se puede concluir que el modelo híbrido planteado resultó ser mejor que las técnicas del mismo modelo híbrido de manera independiente, con una precisión del 97.38%; estos resultados pueden ser comparables con los dataset que contienen más de 20 000 registros, debido a que el data set utilizado contiene similar número de registros (24 420).

7.3 Trabajo futuro

- Adaptar el modelo híbrido propuesto para la solución de problemas de detección de deserción para cajas de ahorros y diversos tipos de créditos, dado que presentan características similares al problema resuelto.
- Plantear un modelo que contenga la técnica Nayve Bayes Tree debido a que en la literatura se observó que la utilización de este para la solución del problema de predicción de deserción de clientes bancarios ofrecen buenos resultados.

8 Referencias

Ali, M., Chernova, T., Newnam, G., Yin, L., Shanks, J., Karpova, T., ... Wilkinson, K. (2014). Stress-dependent proteolytic processing of the actin assembly protein Lsb1 modulates a yeast prion. *J Biol Chem. 289 (40)*, 27625-27639.

Aramburú, C., & Portocarrero, J. (2002). Presentación. *Economía y Sociedad. (46)*, 3-6.

Arredondo, T. (2008). *Árboles de Decisión.* Obtenido de http://profesores.elo.utfsm.cl/~tarredondo/info/soft-comp/Arboles%20de%20Decision.pdf

Athanassopoulos, A. (2000). Customer satisfaction cues to support market segmentation and explain switching behavior. *Journal of Business Research. 47 (3)*, 191-207.

Benoit, D., & Van den, D. (2012). Improving customer retention in financial services using kinship network information. *Expert Systems with Applications. 39 (13)*, 11435-11442.

Chandar, M., Laha, A., & Krishna, P. (2006). Modeling churn behavior of bank customers using predictive data mining techniques. *National Conference on Soft Computing Techniques for Engineering Applications (SCT-2006)*. Rourkela.

Domínguez, E. (s.f.). *Funciones en el modelo de Neurona Artificial.* Obtenido de Introducción a las Redes Neuronales: http://www.redes-neuronales.com.es/tutorial-redes-neuronales/funciones-de-las-neuronas-artificiales.htm

Farquad, M., Ravi, V., & Bapi, R. (2012). Analytical CRM in banking and finance using SVM. *International Journal of Electronic Customer Relationship Management. 6 (1)*, 48-73.

Farquad, M., Ravi, V., & Bapi, R. (2014). Churn prediction using comprehensible support vector machine: An analytical CRM application. *Applied Soft Computing. 19*, 31-40.

Gür, Ö., & Arıtürk, U. (2014). Dynamic churn prediction framework with more effective use of rare event data: The case of private banking. *Expert Systems with Applications. 41 (17)*, 7889-7903.

Jing, Z., & Xing, D. (2008). Bank Customer Churn Prediction Based on Support Vector Machine: Taking a Commercial Bank's VIP Customer Churn as the Example. *Wireless*

Communications, Networking and Mobile Computing, 2008. WiCOM '08. 4th International Conference. Dalian: IEEE.

Keramati, A., Ghaneei, H., & Mohammad, S. (2016). Developing a prediction model for customer churn from electronic banking services using data mining. *Financial Innovation. 2 (10)*, 1-13.

Kitchenham, B., Charters, S., Budgen, D., Brereton, P., Turner, M., Linkman, S., . . . Visaggio, G. (2007). *Guidelines for performing Systematic Literature Reviews in Software Engineering.* Durham: University of Durham.

Kumar, S., & Chandrakala, D. (2016). A Survey on Customer Churn Prediction using Machine Learning Techniques. *International Journal of Computer Applications. 154 (10)*, 13-16 .

Kuncheva, L. (2004). *Combining pattern classifiers methods and algorithms.* Nueva York: John Wiley & Sons, Inc.

León, J., & Jopen, G. (2011). *La heterogeneidad del microcrédito en el sector financiero regulado peruano.* Lima: Pontificia Universidad Católica del Perú.

López, M. C., López, M., & Martínez, S. (2016). A stochastic comparison of customer classifiers with an application to customer attrition in commercial banking. *Scandinavian Actuarial Journal. (7)*, 606-627.

Maldonado, S. (2015). Churn prediction via support vector classification: An empirical comparison. *Intelligent Data Analysis. 19 (s1)*, 135-147.

Marín, J. (2012). *Introducción a las redes neuronales aplicadas.* Madrid: Universidad Carlos III de Madrid.

Miranda, J. (2006). *Modelo de predicción de fugas voluntarias para una institución financiera utilizando Support Vector Machines [Tesis de Maestría]*. Santiago: Universidad de Chile.

Nie, G., Rowe, W., Zhang, L., Tian, Y., & Shi, Y. (2011). Credit card churn forecasting by logistic regression and decision tree. *Expert Systems with Applications. 38 (12)*, 15273-15285.

Portocarrero, F., Trivelli, C., & Alvarado, J. (2002). *Microcrédito en el Perú: quiénes piden, quiénes dan.* Lima: Consorcio de Investigación Económica y Social.

Quispe, Z., Leon, D., & Contreras, A. (2011). Exitoso desarrollo de las microfinanzas en el Perú. *Revista Moneda Microfinanzas. (151)*, 13-18.

SBS. (s.f.). *Inicio.* Obtenido de http://www.sbs.gob.pe/

Sin, C., Hshiung, G., & Chieh, Y. (2011). Combined rough set theory and flow network graph to predict customer churn in credit card accounts. *Expert Systems with Applications. 38 (1)*, 8-15.

Sundarkumar, G., & Ravi, V. (2015). A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance. *Engineering Applications of Artificial Intelligence. 37*, 368-377.

Tang, L., Thomas, L., Fletcher, M., Pan, J., & Marshall, A. (2014). Assessing the impact of derived behavior information on customer attrition in the financial service industry. *European Journal of Operational Research. 236 (2)*, 624-633.

Trujillano, J., Badía, M., March, J., Rodríguez, C., Serviá, L., & Sorribas, A. (2005). Redes neuronales artificiales en Medicina Intensiva. Ejemplo de aplicación con las variables del MPM II. *Medicina Intensiva. 29 (1)*, 13-20.

Van den, D., & Lariviere, B. (2004). Customer attrition analysis for financial services using proportional hazard models. *European Journal of Operational Research. 157 (1)*, 196-217.

Wang, N., & Niu, D. (2009). Credit Card Customer Churn Prediction Based on the RST and LS-SVM. *Service Systems and Service Management, 2009. 6th International Conference on.* Xiamen.

WebMining Consultores. (2011). *KDD: Proceso de Extracción de conocimiento.* Obtenido de http://www.webmining.cl/2011/01/proceso-de-extraccion-de-conocimiento/

WZL. (s.f.). *Karriere am WZL.* Obtenido de Production Logistics: http://www.wzl.rwth-aachen.de/en/f27f422f5350ada9c1257dfd004b21a3.htm

Xanadu Linux. (2017). *El perceptrón y perceptrón multicapa ¿Qué es y con que se come?* Obtenido de El Blog de Sinfallas: https://sinfallas.wordpress.com/2017/11/14/el-perceptron-y-perceptron-multicapa-que-es-y-con-que-se-come/

Xiao, J., Jiang, X., He, C., & Teng, G. (2016). Churn Prediction in Customer Relationship management via GMDH. *IEEE Intelligent Systems. 31 (2)*, 37-44.

Xie, Y., & Li, X. (2008). Churn Prediction with Linear Discriminant Boosting Algorithm. *Machine Learning and Cybernetics, 2008 International Conference.* Kunming.

Xie, Y., Li, X., Ngai, E., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications. 36 (3)*, 5445-5449.

Yepes, V. (2017). *¿Qué es y para qué sirve una red neuronal artificial?* Obtenido de El blog de Víctor Yepes: http://victoryepes.blogs.upv.es/2017/01/07/que-es-y-para-que-sirve-una-red-neuronal-artificial/

Zhao, X., Shi, Y., Lee, J., Kee, H., & Lee, H. (2014). Customer Churn Prediction Based on Feature Clustering and Nonparallel Support Vector Machine. *International Journal of Information Technology & Decision Making. 13 (5)*, 1013-1027.

Zhu, B., Xiao, J., & He, C. (2014). A Balanced Transfer Learning Model for Customer Churn Prediction. *Eighth International Conference on Management Science and Engineering Management.* Berlin: Springer.

T. Vafeiadisa, K. I. Diamantaras, & G. Sarigiannidis (2015). A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory

A. De Caignya, K. Coussementa, & Koen W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. European Journal of Operational Research 269, 760–772

R. Porter,Linda A. Miner, & G. Miner (2018). Using Customer Churn Data to Develop and Select a Best Predictive Model for Client Defection Using STATISTICA Data Miner 13 64-bit for Windows 10. Handbook of Statistical Analysis and Data Mining Applications

K. Ravi, V. Ravi, & P. Sree Rama Krishna (2017). Fuzzy Formal Concept Analysis based Opinion Mining for CRM in Financial Services. Applied Soft Computing

F. Shirazia, & M. Mohammadi (2018). A big data analytics model for customer churn prediction in the retiree segment. International Journal of Information Management.